**1 i 2 Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? Integració i selecció de les dades d’interès a analitzar.**

El dataset escollit es un conjunt d’informació sobre els Pokemon. Els pokemon son un conjunt d’animals o personatges d’una de les franquicies més importants de tot el món. Aquesta franquicia és un joc competitiu creat al Japó per GameFreak sota la tutela de Nintendo, amb l’objectiu de superar diversos enemics finals i captura el major nombre possible de Pokemons.

Amb el pas dels anys, aquesta franquícia ha evolucionat fins ocupar un lloc important en el panorama dels esports electronics, doncs ofereix una gran varietat de Pokémon diferents (900), tots amb les seves estadístiques pròpies i habilitats en molts casos úniques, dintre els quals s’han d’escollir 6 per generar un equip amb les millors sinergies possible.

En vista d’aquest fet, hi ha un volum de dades certament interesant que ens pot servir per tractar de cercar quines classes són més fortes, o si tant sol hi ha diferencia entre elles.

Centrant-nos més amb el dataset podem trovar un total de 50 columnes i 1028 files. De les 54 columnes trobem 13 categóriques i 37 numériques, i és pot trobar al següent enllaç:

<https://www.kaggle.com/mariotormo/complete-pokemon-dataset-updated-090420?select=pokedex_%28Update_05.20%29.csv>

Tot seguit explicarem les variables categóriques més importants:

**Nom**: Aquesta variable ens informa del nom del Pokémon en questió. És una variable important ja que és el nom típic utilitzat per nombrar el Pokémon en questió

**Especie**: Dins del món Pokémon la especie es molt important ja que ens donarà molta información al respecte del Pokémon que estem tractant. Ens podem trovar diversos tipus de Pokémon, com per exemple aquatics, de muntanya etc. La especie ens donarà una primera impessió del tipus de Pokémon que estem tractant.

**Tipus 1** i **tipus 2**: Aquestes variables ja ens donen més informació sobre el Pokémon, ja que per a cada especie hi ha tipus de Pokemons. Per exemple, un Pokémon marí pot ser del tipus medusa o del tipus peix.

**Abilities:** Aquest apartat explica quines son les habilitats més importants del Pokémon. Per exemple, depén del tipus i l’especie d’un Pokémon tindrà unes habilitats o altres.

A continuació, comentarem les numériques:

**Total poins:** és la suma de punts d’atac, defensa, velocitat..Et dona una idea del nivel del Pokémon i quina importancia pot tenir.

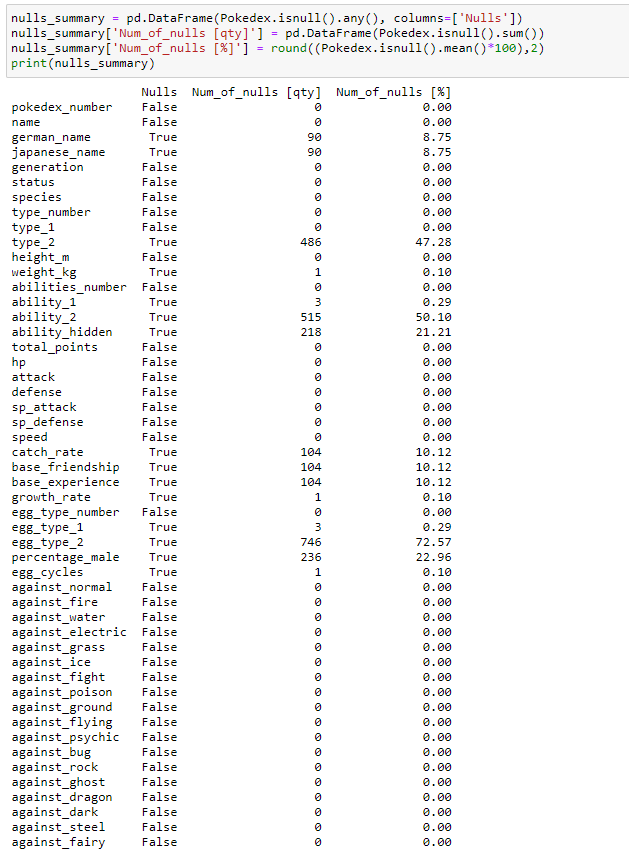
**Estadístiques:** Conjunt de variables numériques que donen diversa información sobre el valor, el nivel i la forma d’un Pokémon.

**Neteja de les dades.**

**3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?**

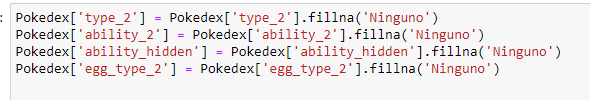
Un cop vistes la informació més rellevant sobre el nostre dataset, és moment de veure fins quin punt ens és útil la informació de les 50 variables, i també veure si és necessari fer algun tractament previ de les dades.

Per assolir aquests objectiu, en primer lloc el hem fet és analitzar la presencia de valors nuls i zeros a les dades. Per fer-ho hem utilitzat els següents codis a python:

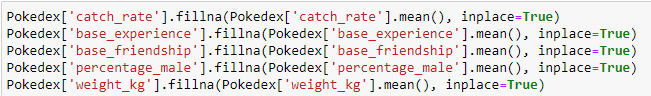


Com podem veure, pel que fa als valors nuls (elements buits), veiem com algunes variables tenen una forta presencia de valors nuls, com ara el grup ou secundari (egg\_type\_2) o el tipus secundari del Pokémon. Davant d’aquesta situació ens trobem amb dues alternatives, eliminar les columnes i/o files amb amb NA o bé buscar un mètode d’imputació de valors adient.

Aquests cop, hem optat per tractar d’entendre si són valors nuls explicables o no, i sempre que fos possible fer una imputació, doncs esborrar les files seria obviar gran part d’informació. Analitzant les variables, hem vist com els valors absents en type\_2, ability\_2, hidden\_ability i egg\_type\_2 eren nuls donat que al ser atributs addicionals no tots els Pokémon en tenen, i per tant, s’ha optat per recodificar els valors nuls com ‘Ninguno’:

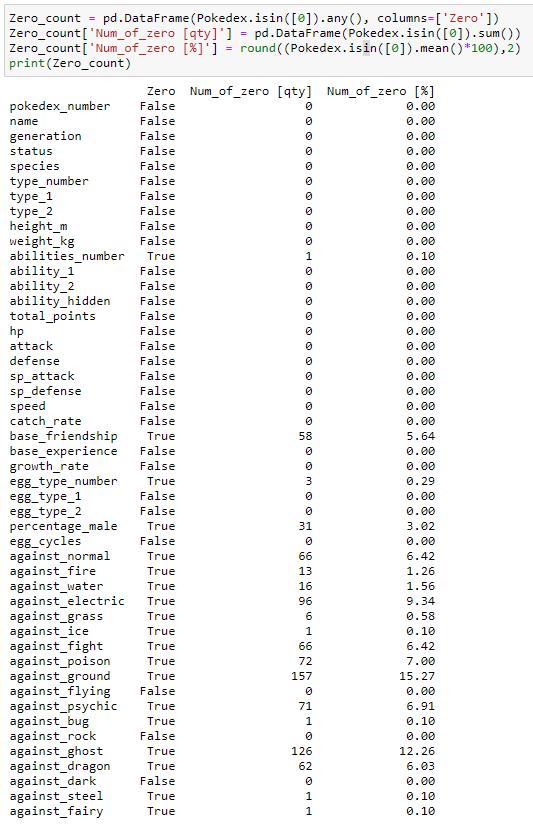


D’altre banda, els casos de catch\_rate, base\_friendship i percentatge\_male s’han solucionat imputant el valor mitja de la variable:



Finalment, s’ha optat per eliminar les columnes amb el nom alemany i japonès, doncs a més de tenir valors buits no aportaven cap informació rellevant. Així mateix, s’han eliminat les files restants amb NA.

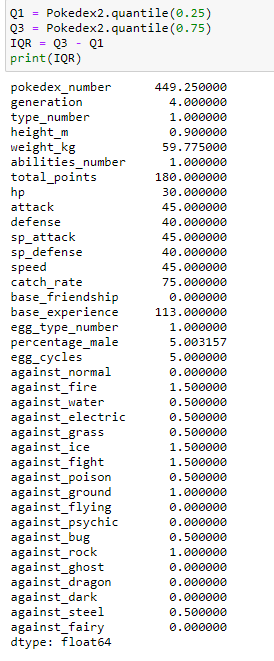
Pel que fa referencia a les observacions amb valor 0, s’han localitzat les següents:



Aquest cop, no ha estat necessari eliminar cap registre ni modificar els seu valor, doncs tots eren valors coherents amb els esperats per les variables mostrades.

**3.2. Identificació i tractament de valors extrems.**

Pel que fa als valors outliers, hem calculat el rang interqualitic per veure quins valors són el límit que marquen si és valor atípic o no:



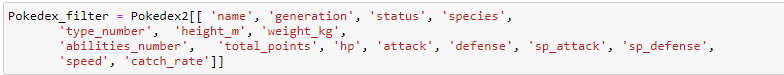
No obstant, després de veure en detall els casos en els que succeïa, s’ha decidit conservar aquestes observacions, ja que eliminar-los reduiria en un al percentatge la base.

**4. Anàlisi de les dades.**

**4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels**

**anàlisis a aplicar).**

Tot seguit, és seleccionaran les variables que farem servir per realitzar les proves estadístiques, per realitzar les comparacions que ens resultin interesants. No obstant, finalment no totes les variables i grups escollits seran utilitzats en aquests apartat.

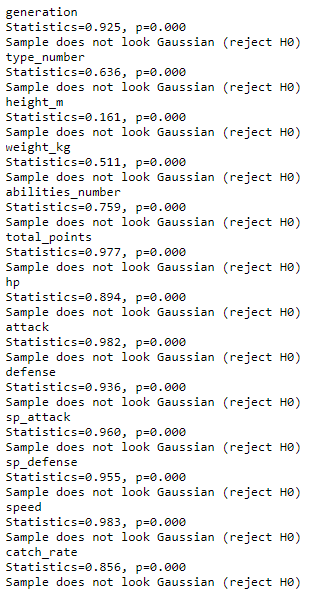


Així mateix, les variables status i species seran dummificades en variables auxiliars en funció dels valors de les seves categories.

**4.2. Comprovació de la normalitat i homogeneïtat de la variància.**

Per tal de comprovar la normalitat i homogeneïtat de la variància s’ha aplicat el test de Shapiro-Wilk:





Com podem veure, la majoria de variables refusen la hipòtesi nul·la de normalitat. És a dir, segons el test de Shapiro Wilk no segueixen una distribució normal. Només la variable pes segueix una distribució normal

**4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d’anàlisi diferents.**

**Model regressió lineal:**

En aquest model, s’intenta fer una regressió lineal per predir el ‘catch\_rate’. Aquesta variable simbolitza la dificultat d’un Pokémon al ser capturat. Un Pokémon quan es vol atrapar pot tenir una dificultat major o menor per ser atrapat. Es vol veure quines son les variables que afecten a la dificultat i si hi ha alguna manera de predir-ho de forma correcte.

Les variables utilitzades són: generació, status, pes, alçada, nombre d’abilitats, atac, defensa, sp\_atac, sp\_defensa.

S’utilitza variables dummies status y specie per veure el comportament. El R^2 està per sobre de 0.87 i es consegueixen bon resultats

Eliminant del model aquestes 2 varibles es torna a computar i s’obté un 0.49 per tant s’interpreta que les variabels de specie i status generen gran importància ja que sense elles la variança explicada es redueix considerablement.

|  |  |
| --- | --- |
| **Contribucions** | **Signa** |
| **Recerca prèvia** | ACL i JML |
| **Redacció de les respostes** | ACL i JML |
| **Desenvolupament codi** | ACL i JML |

El codi i el dataset es pot trobar al següent enllaç:

<https://github.com/jordim14/Pokedex>