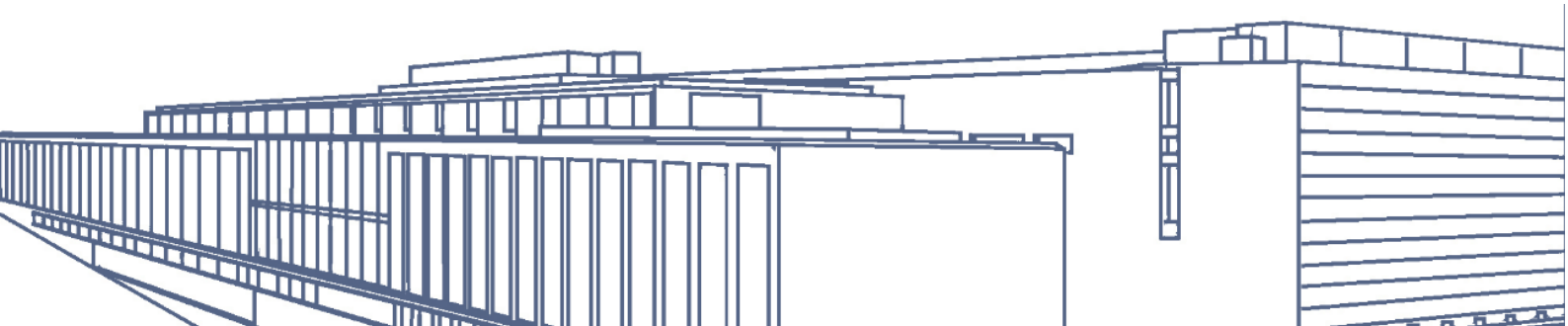
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA
Universitat Rovira i Virgili

Sistemes d'Informació a les Organitzacions

Pràctica 1 (Part 1): Imputació de valors desconeguts: A la recerca de les millors prediccions

16/03/2020

Victor López Romero
Aleix Sancho Pujals



Contingut

1. Introducció

1.1. Resum de la pràctica

En aquesta primera part de la pràctica es demana obtenir informació sobre el conjunt de dades proporcionat, és a dir, fer una anàlisi exploratòria. Així mateix, per a realitzar l'anàlisi es pot usar qualsevol plataforma d'anàlisi estadística que es trobi a disposició (e.g., R, SPSS) i qualsevol llenguatge de programació (e.g., Java, C, Python, ...).

Un cop les dades estan organitzades, cal fer-se preguntes sobre la seva estructura, correlacions, dependències, etc.

1.2. Conjunt de dades

El conjunt de dades que hem d'analitzar són numèriques discretes, degut que són valors numèrics amb coma flotant de dos decimals.

El conjunt de dades es proporciona en format csv (format numèric separat per ";"). Cada fila correspon al perfil d'un usuari (**vector usuari**) i cada columna correspon a les valoracions de tots els usuaris sobre un ítem (**vector ítem**). Base de dades de recomanacions de restaurants (4M valoracions reals).

- Total usuaris **73,421**
- Total ítems **100**
- Rang de valoracions **[-10, 10] Reals amb 2 decimals**
- Densitat **55.8%**
- **Valor buit = 99 (les caselles buides tenen un valor = 99)**

2. Treball realitzat

2.1. Base de dades

Per a poder analitzar les dades, hem decidit utilitzar un gestor de base de dades, en concret PostgreSQL. Per a gestionar la base de dades i tractar les dades hem decidit utilitzar el llenguatge Python. Hem estructurat la base de dades de la següent manera:

- Una taula on introduïm tots els usuaris, estructurada de la següent manera:
    - Id de l'usuari, not null i unic, que a la vegada serà clau primària.
- Una taula on introduïm tots els restaurants, estructurada de la següent manera:
    - Id del restaurant, not null i unic, que a la vegada serà clau primària.
- Una taula de relacions entre usuaris i restaurants on guardarem les puntuacions, estructurada de la següent manera:
    - Id de l'usuari, not null i clau forana a l'id de l'usuari de la taula dels usuaris.
    - Id del restaurant, not null i clau forana a l'id del restaurant de la taula dels usuaris.
    - Puntuació, not null.

Una vegada creada la taula en el cas de la dels usuaris i els restaurants s'ha omplert amb tots els Ids que es troben en el fitxer csv. Però en el cas de la taula de relacions s'han descartat totes aquelles entrades que tenien un 99 de puntuació, és a dir les dades que són nul·les.

2.2. Dades bàsiques extretes de la base de dades

Una vegada totes les dades han estat introduïdes en la base de dades, hem llegit aquestes dades per extreure informació bàsica.

Hem detectat que hi ha un total de **4.096.360** de puntuacions en la base de dades, quan realment en el csv hi ha **73.421.000** relacions.

La puntuació més alta que s'ha produït és un **10** i la més baixa és **-9,95**.

Hem observat que el nombre més petit de visites que ha realitzat un usuari és de **14** vegades, i això ho han fet un nombre de **27** usuaris. Per altra banda, en l'altre extrem s'han realitzat **100** visites, és a dir que han valorat tots els restaurants i això ho han fet **86** usuaris diferents.

2.3. Anàlisi de les dades

2.3.1. Recompte de puntuacions en intervals

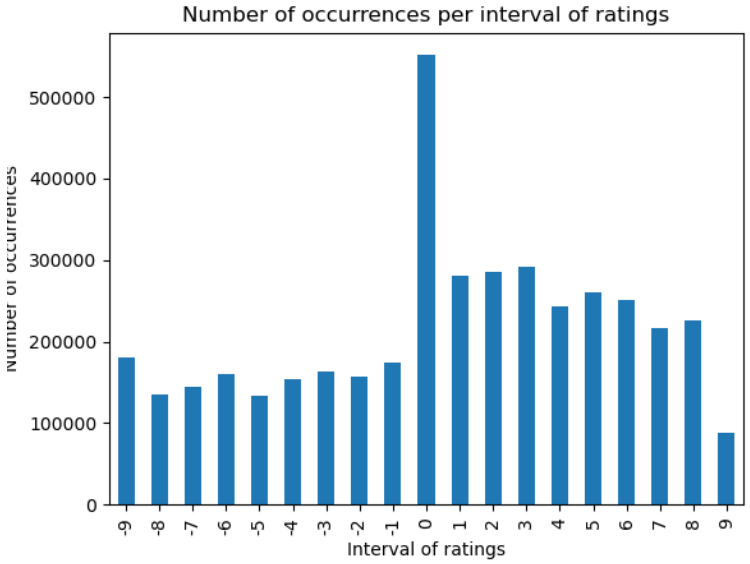


Figura 1

Puntuació	Vegades que apareix
-10	96142
-9	149580
-8	139251
-7	152440
-6	145807
-5	143189
-4	165200
-3	154172
-2	167809
-1	211028
0	299134
1	261463
2	286576
3	290189
4	273478
5	253191
6	259993
7	230491
8	212657
9	204570

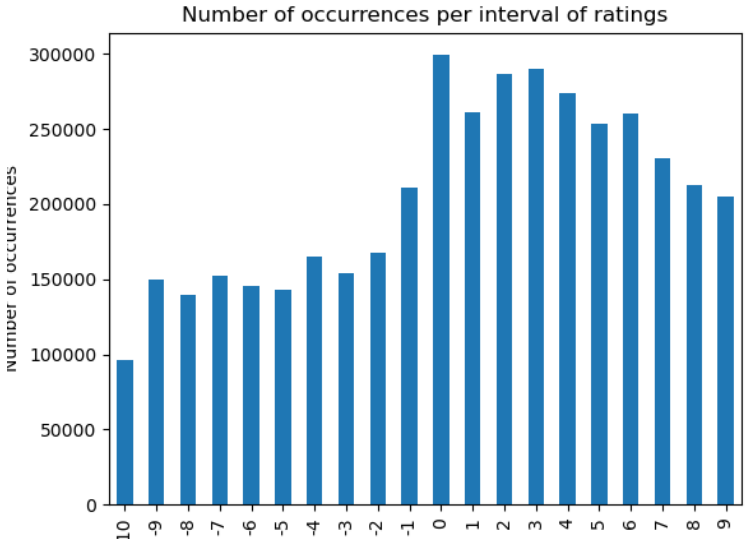


Figura 2

En aquesta gràfica es mostren les dades que hem extret a partir de les diferents puntuacions que han realitzat els usuaris, com tenim moltes dades diferents, hem decidit agrupar aquestes dades en diferents intervals enters ([-x.99,x.99]) a les dades analitzades. Un cop agrupades en aquests intervals hem fet un recompte del nombre de puntuacions que cauen en cada interval, en la Figura 1 es pot veure que la majoria de puntuacions cauen en l'interval [-0.99, 0.00] això és degut que s'ha utilitzat un sistema de truncament de les dades, ja que en la Figura 2 s'han arrodonit les dades i es mostra una major distribució d'aquestes puntuacions. Cal dir que les puntuacions al voltant del 10 s'han agrupat al 9 perquè eren molt poques.

Com a conclusió final d'aquestes gràfiques podem concloure que la majoria de puntuacions estan en l'interval [-0.99, 0.99], però que la majoria d'elles tendeixen a anar a l'alça. També que entre totes les puntuacions s'han realitzat més puntuacions positives que negatives.

2.3.2. Probabilitat de cada puntuació

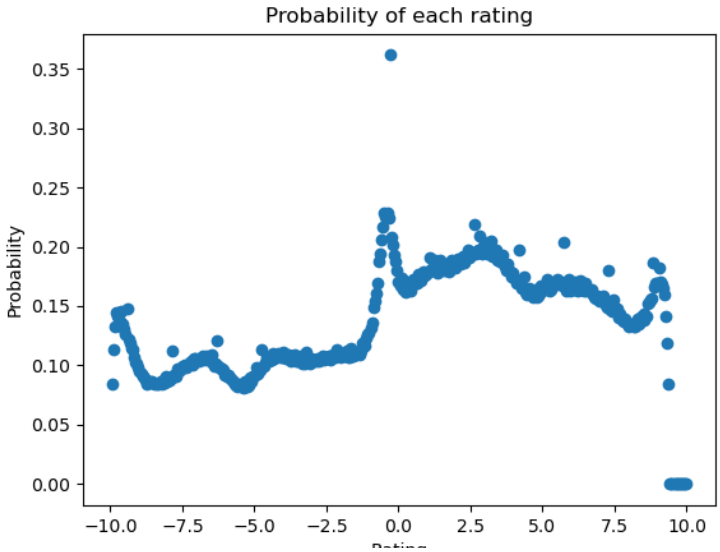


Figura 3

A la gràfica anterior es pot observar les probabilitats que té una puntuació de ser escollida. Com es pot observar les puntuacions menys extremes són les més probables de sortir. Aquesta gràfica també es complementa amb les anteriors on ens diu que **la puntuació més probable és una que estigui sobre el 0 i amb una mica menys de probabilitat qualsevol puntuació positiva que no sigui molt extrema**. Tot i això no podem extreure molta informació d'aquesta gràfica, ja que les probabilitats són baixes.

2.3.3. Recompte de les mitjanes de puntuacions dels usuaris en intervals

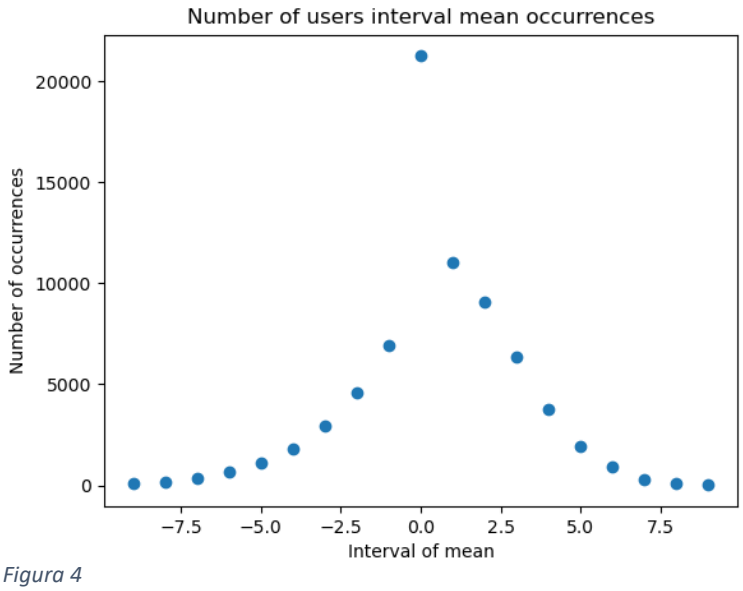


Figura 4

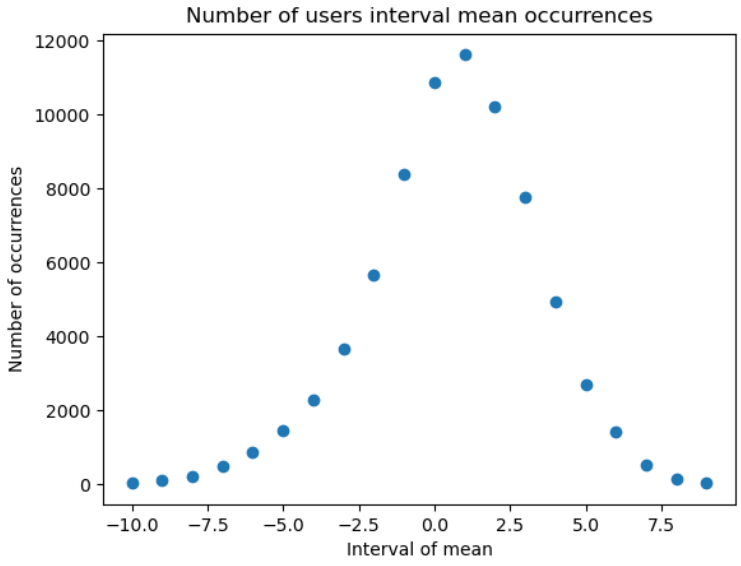


Figura 5

mean	count
-10	26
-9	100
-8	223
-7	505
-6	854
-5	1453
-4	2283
-3	3663
-2	5663
-1	8399
0	10868
1	11618
2	10205
3	7781
4	4924
5	2705
6	1410
7	536
8	159
9	46

A la gràfica anterior es pot observar el nombre d'aparicions dels diferents intervals enters ([-x.99,x.99]) de les mitjanes de puntuacions de cada usuari a les dades analitzades. Es pot observar fàcilment com **l'interval [-0.99, 0.99] és el més trobat a les dades**. És una prova similar a l'anterior però treballant amb les mitjanes. De la mateixa manera que abans tenim la Figura 4 realitzada amb truncament de dades i la Figura 5 amb arrodoniment.

2.3.4. Positivisme dels usuaris

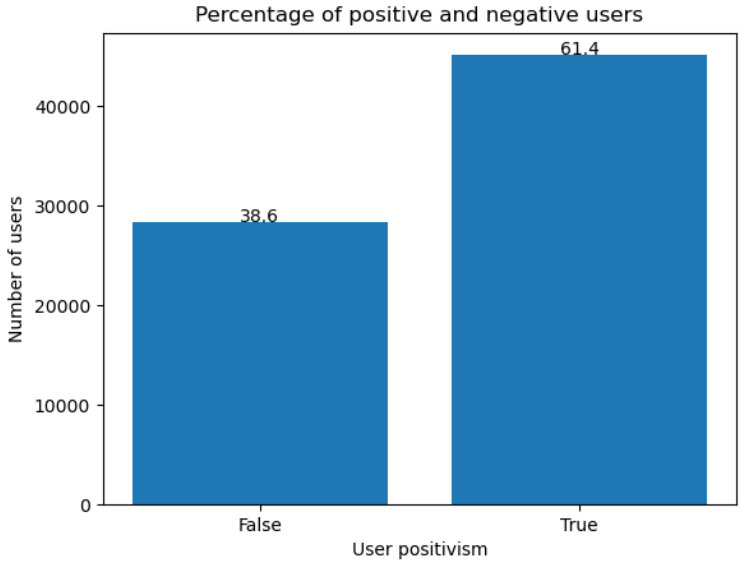


Figura 6

A la gràfica anterior es pot observar el nombre d'aparicions de les mitjanes de les puntuacions dels usuaris dividits en dues parts, els valors positius i els negatius. Com es pot observar a **la gràfica la majoria dels usuaris tenen mitjanes positives**. El valor False fa referència a un usuari

negatiu i el True a un positiu, ens hem basat en el fet que buscar el nombre de vegades que un usuari puntua per sobre de 0 o per damunt d'aquesta puntuació.

2.3.5. Positivisme dels restaurants

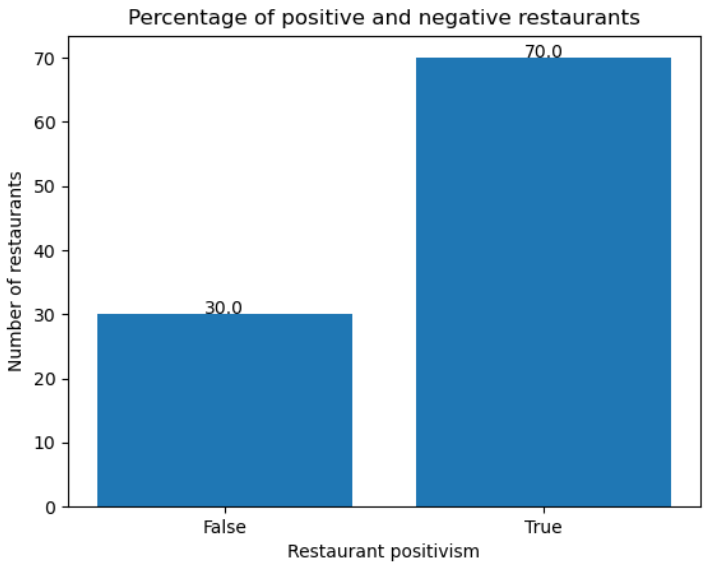


Figura 7

A la gràfica anterior es pot observar el nombre d'aparicions de les mitjanes de les puntuacions dels restaurants dividits en dues parts, els valors positius i els negatius. Com es pot observar a la gràfica **la majoria dels usuaris tenen mitjanes positives.**

El valor False fa referència a un restaurant negatiu i el True a un positiu, ens hem basat en el fet que buscar el nombre de vegades que un usuari puntua per sobre de 0 o per damunt d'aquesta puntuació.

2.3.6. Nombre de visites dels restaurants per la seva mitjana

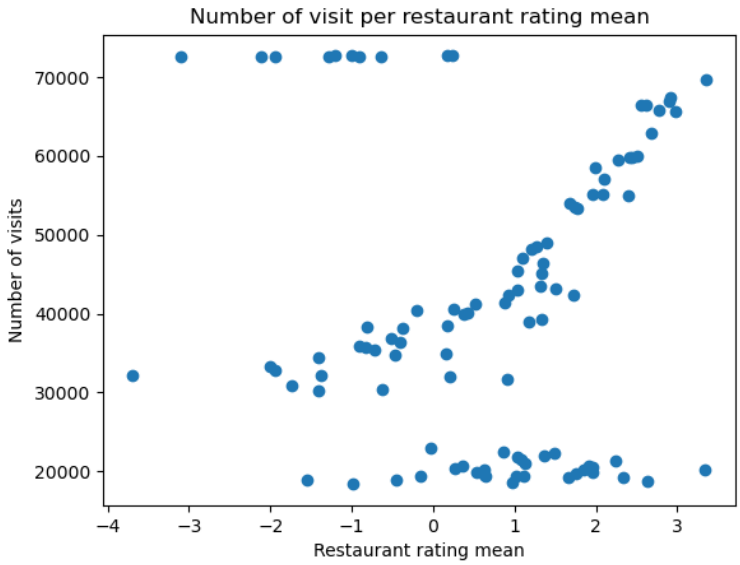


Figura 8

En aquesta gràfica volem analitzar quina afluència té cada restaurant respecte la seva puntuació mitjana, és a dir si depèn aquesta afluència de la puntuació que ha obtingut com a mitjana. Podem veure en la Figura 8 que **es diferencien 3 grans grups** que es podrien analitzar de manera separada. El primer d'ells es tracta dels **restaurants que tenen una puntuació mitjana entre -4 i 0 i tenen una gran quantitat de visites**, en aquest grup hi pertanyen 10 restaurants. Un altre grup seria just el cas contrari, on els **restaurants tenen una mitjana de puntuació més elevada, entre -1 i 4 però tenen una afluència de gent molt baixa**, hi **agrupem 30 restaurants**. Per tant **la resta de restaurants, 60, segueixen una tendència lineal, on podem concloure que com més afluència tenen, reben una puntuació mitjana més positiva.**

2.3.7. Probabilitat que una mateixa puntuació la faci diferents usuaris

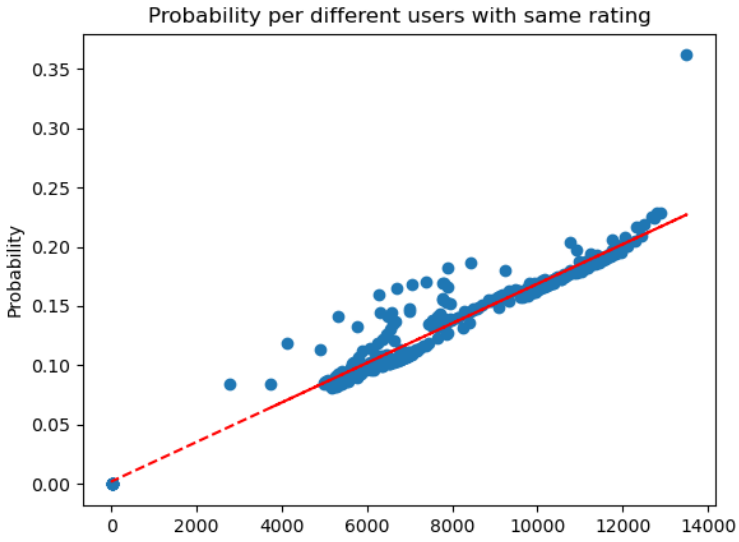


Figura 9

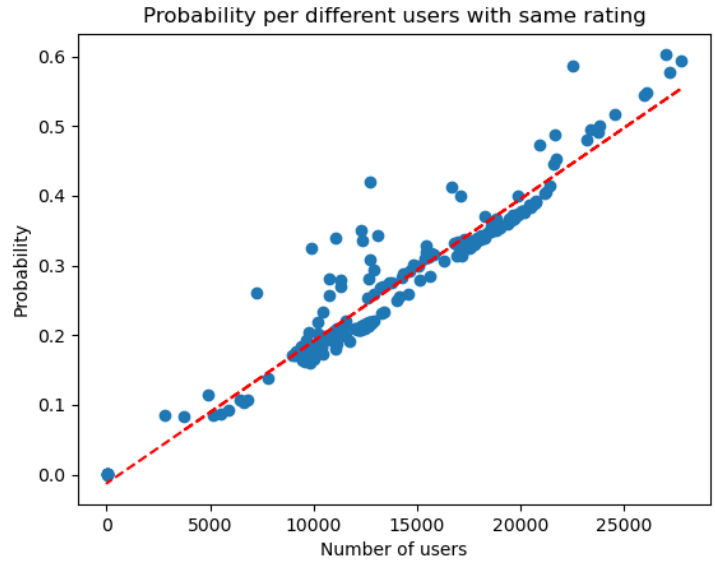


Figura 10

Per poder fer aquesta anàlisi el primer que vam decidir fer va ser disminuir el nombre de puntuacions per poder agrupar els usuaris en grups més grans, d'aquesta manera les puntuacions s'han truncat/arrodonit en la Figura 9 i 10 respectivament a 1 decimal. Una vegada fet això, s'ha afegit la probabilitat que hi ha que es repeteixi una puntuació, calculada anteriorment i finalment hem contat el nombre d'usuaris diferents que han puntuat amb aquesta puntuació. Com s'ha esmentat en la gràfica mostrem la relació que segueix el nombre d'usuaris diferents que han puntuat amb certa nota i la probabilitat que té aquesta d'aparèixer. Els resultats obtinguts són els esperats i és que **com més usuaris diferents puntuen amb certa nota, és més probable que aquesta es repeteixi, per tant ens assegurem que no hi ha cap usuari que faci servir moltes vegades una mateixa nota** i això faci analitzar malament les gràfiques anteriors.

2.3.8. Comparació entre la puntuació mitja dels usuaris i la seva moda (I)

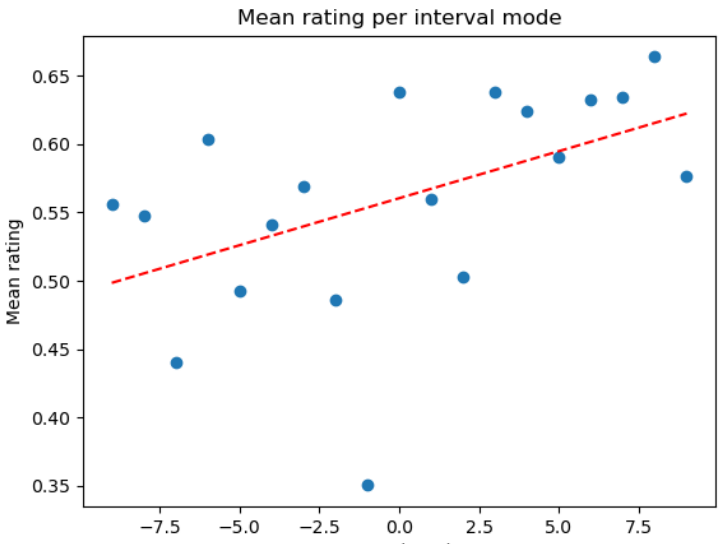


Figura 11

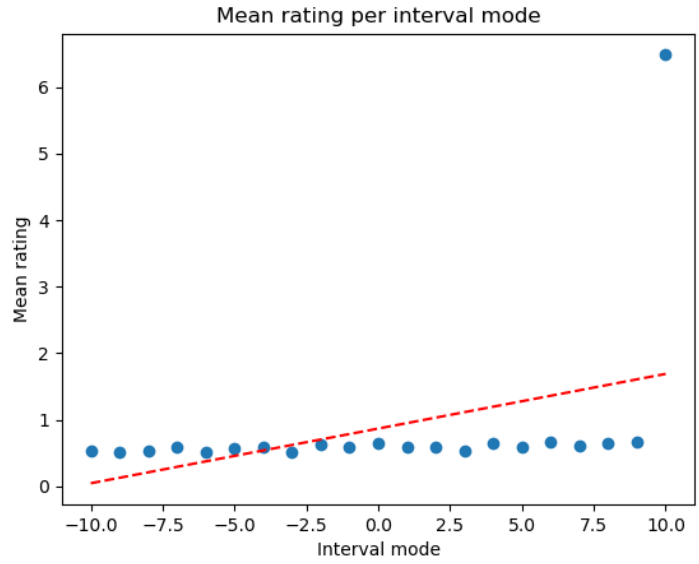


Figura 12

Per començar a analitzar la gràfica s'ha de comentar que per poder treballar amb la moda dels usuaris, aquesta s'ha truncat en la Figura 11 i s'ha arrodonit en la Figura 12 a un nombre enter. Per agrupar la moda amb la mitjana de cada usuari hem decidit fer la mitjana de la mitjana de cada usuari que comparteixin la mateixa moda, de tal manera que podem concloure que **la moda és bastant semblant a la mitjana, és a dir que les puntuacions dels usuaris augmenten a mesura que també augmenta la moda.**

2.3.9. Comparació entre la puntuació mitja dels usuaris i la seva moda (II)

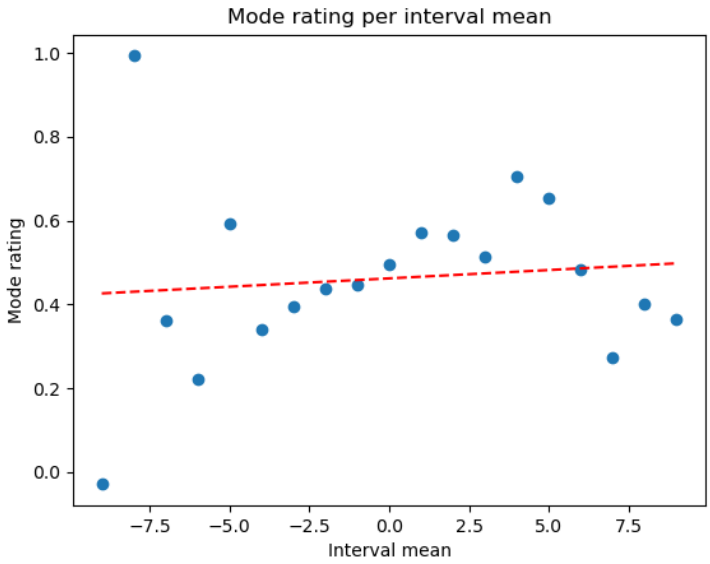


Figura 13

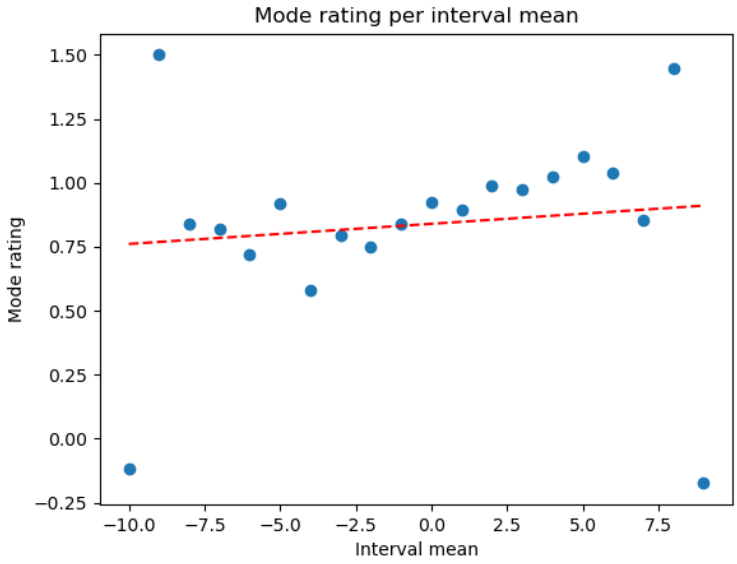


Figura 14

En aquest estudi fem una anàlisi similar a l'anterior, partim amb les mateixes dades, però les analitzem de manera una mica diferent. Cal dir que de la mateixa manera la primera figura ho fem amb truncament de dades i a la segona amb arrodoniment d'aquestes. Per tal d'obtenir aquest gràfic el que farem és reduir el nombre de mitjanes agrupant-les en grups d'enters, d'aquesta manera es redueix a 20 dades i posteriorment les podem comprar amb les modes. Per tant el que farem serà agrupar totes les mitjanes que tinguin el mateix valor i farem la mitjana de les modes corresponents. Per tant, extraiem que **a mesura que la mitjana de puntuacions incrementa també ho fa la moda que pertanyia a aquestes puntuacions mitjanes**.

2.3.10. Desviació estàndard poblacional respecte la moda dels usuaris

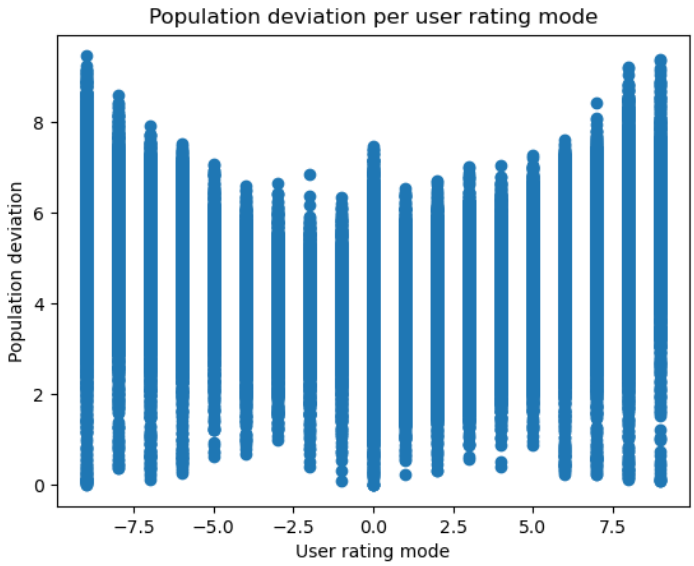


Figura 15

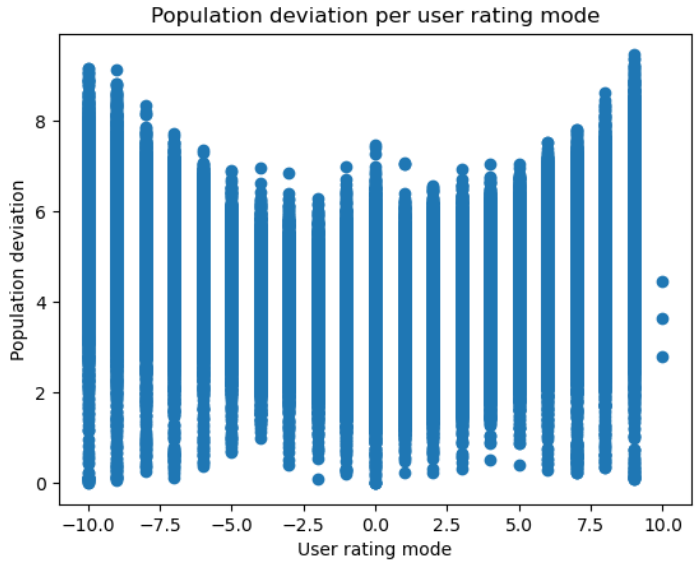


Figura 16

Per fer l'anàlisi d'aquesta gràfica hem utilitzat els valors de la moda aproximada a 1 enter, en la Figura 15 per truncament de dades i en la Figura 16 per arrodoniment. Aquests valors els comparem amb la desviació estàndard poblacional de les puntuacions dels usuaris i podem extreure que **hi ha més desviació de les puntuacions a mesura que la moda d'aquestes tendeixen a 0, o el que seria el mateix, que en les modes extremes hi ha més desviació.**

2.3.11. Desviació estàndard poblacional dels restaurants respecte la seva mitjana

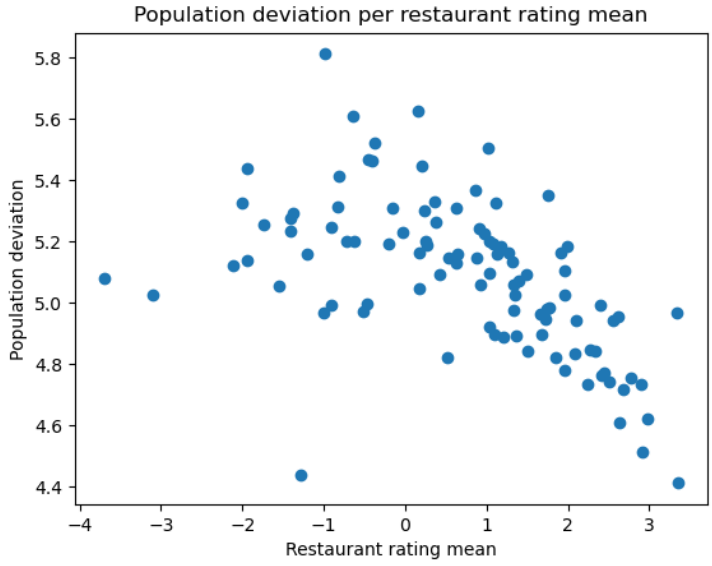


Figura 17

En aquesta figura mostrem una relació semblant a l'anterior però en comptes d'utilitzar la moda dels usuaris, en aquest cas farem servir la mitjana de les puntuacions però agrupades per restaurants. Podem veure que hi ha una relació entre la desviació estàndard poblacional dels restaurants i la mitjana de les puntuacions que han rebut aquests. **A mesura que la puntuació mitjana dels restaurants augmenta, disminueix la desviació que aquests pateixen.**

2.3.12. Desviació estàndard poblacional dels usuaris respecte la seva mitjana

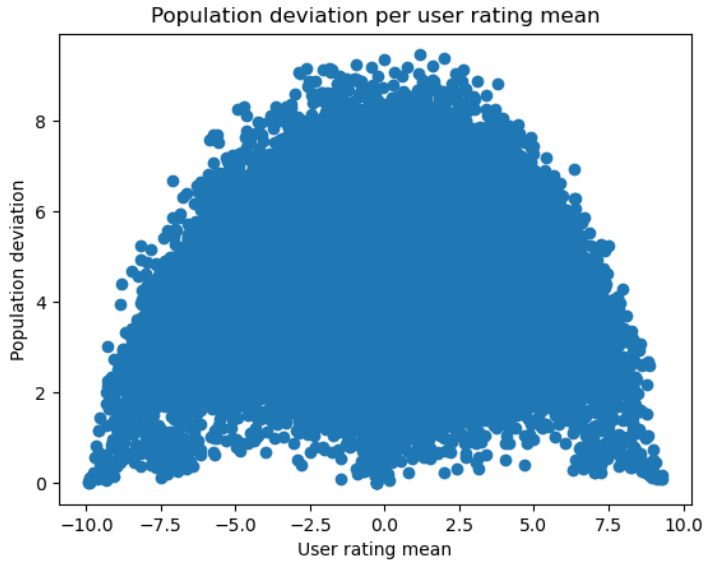


Figura 18

Per obtenir aquesta relació s'han utilitzat la mitjana de cada usuari i la desviació estàndard poblacional d'aquests. Podem veure que **com més neutres són les puntuacions mitjanes, més desviació tenen, per tant en els extrems sembla que les puntuacions són més estables**.

2.3.13. Mitjanes de les puntuacions dels restaurants en intervals

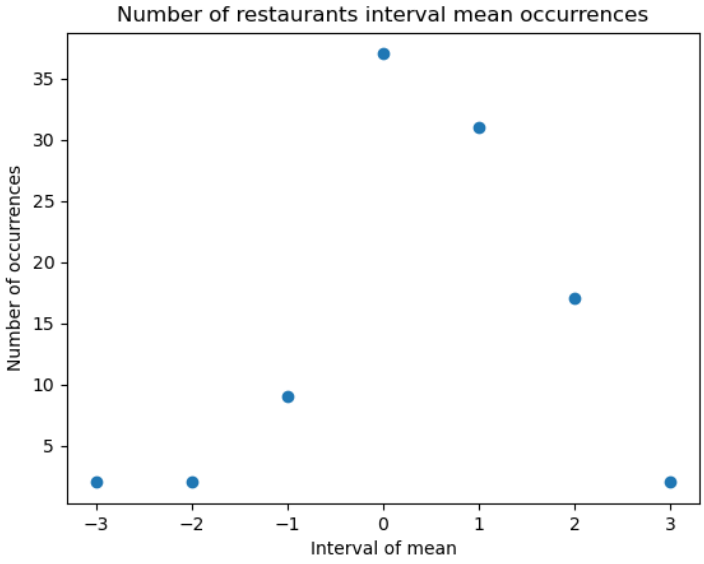


Figura 19

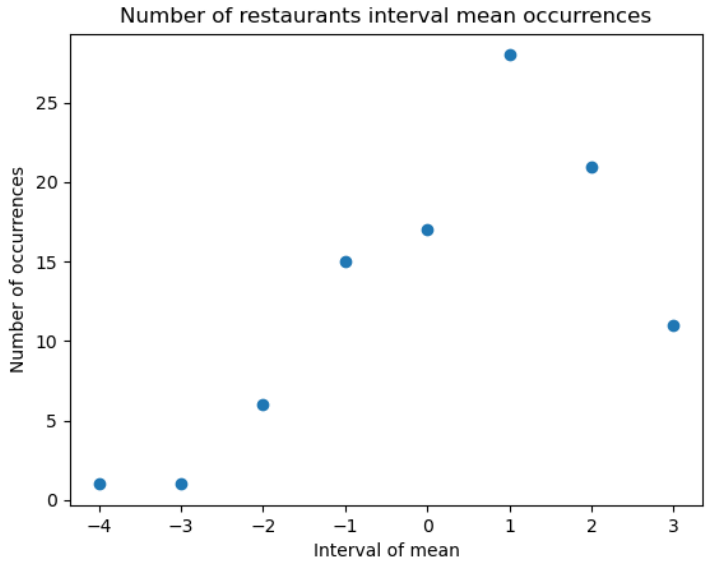


Figura 20

En aquesta anàlisi hem agrupat la mitjana de les puntuacions dels restaurants en intervals d'enters per reduir les dades a tractar i les hem comparat amb el nombre de vegades que una puntuació cau en cada interval. En la Figura 19 s'han fet els intervals per truncament i en la Figura 20 per arrodoniment i podem extreure, com s'ha vist anteriorment, que **en les puntuacions que van de -1 a 1 és on cauen la majoria de puntuacions mitjanes dels restaurants**.

2.3.14. Nombre d'afluència en els restaurants per la seva mitjana

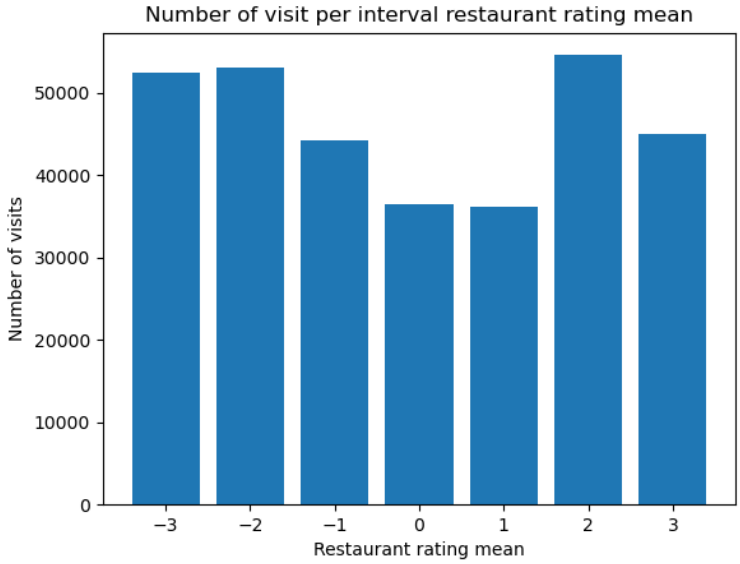


Figura 21

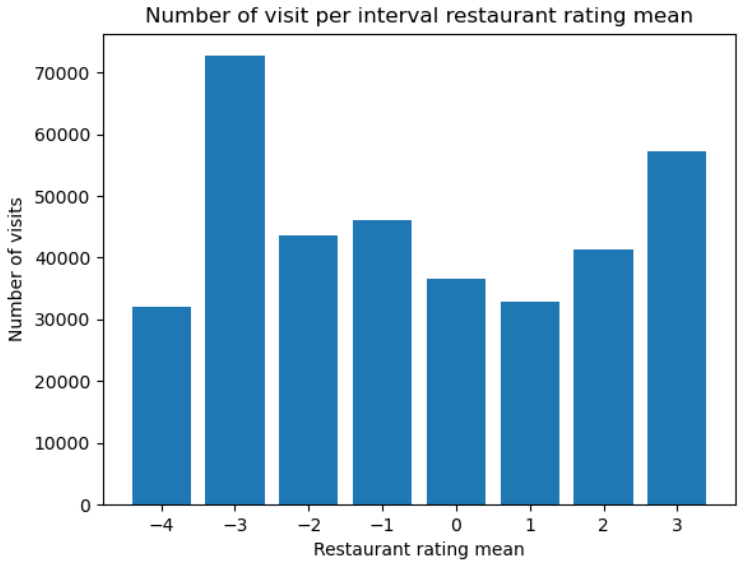


Figura 22

Com ha passat en casos anteriors, hem arrodonit la mitjana dels restaurants a un nombre enter per poder analitzar d'una manera més senzilla les dades, tot i haver-hi una petita pèrdua d'informació. En la Figura 21 s'ha fet amb truncament i en la 22 per arrodoniment. Podem observar gràcies al segon gràfic que els restaurants més visitats aproximen la seva puntuació als voltants del -3, per tant **com més es visita un restaurant aquest tendeix a obtenir una puntuació negativa. En menys afluència de gent, passa alguna cosa similar amb les puntuacions més positives, i on obtenim menys afluència de gent els restaurants obtenen una puntuació més neutra, entre el 0 i l'1.**

3. Conclusions

Gràcies a l'anàlisi realitzat hem obtingut algunes conclusions de les dades proporcionades.

- El conjunt de dades que hem d'analitzar són numèriques discretes, degut que són valors numèrics amb coma flotant de dos decimals.
- La majoria de puntuacions estan en l'interval [-0.99, 0.99], però que la majoria d'elles tendeixen a anar a l'alça.
- Entre totes les puntuacions s'han realitzat més puntuacions positives que negatives.
- La puntuació més probable es realitza al voltant de 0 i amb una mica menys de probabilitat qualsevol puntuació positiva que no sigui molt extrema.
- Com més usuaris diferents puntuen amb certa nota, és més probable que aquesta es repeteixi.
- La moda es comporta de manera semblant a la mitjana.
- Hi ha més desviació de les puntuacions a mesura que la moda d'aquestes tendeixen a 0.
- Com més neutres són les puntuacions mitjanes, més desviació tenen.
- Com més es visita un restaurant aquest tendeix a obtenir una puntuació negativa.
- Els restaurants amb menys afluència de gent obtenen una puntuació més neutra (entre el 0 i 1).
- Com més visitat és un restaurant, aquest rep una puntuació més extrema.