(پروژه داده کاوی)

گروه: زهرا رستمی – مینا افضلی

فاز دوم) ارزیابی کیفیت داده

feature name	number of records	number of Null values	Accuracy	completness	validity	currentness	consistency
img	64016	0	84.87%	1	0.848725	-	-
title	64016	0	55.10%	1	0.551034	-	-
console	64016	0	91.85%	1	0.918458	-	-
genre	64016	0	69.14%	1	0.691421	-	-
publisher	64016	0	96.75%	1	0.967477	-	-
developer	64016	17	91.58%	0.999734	0.915802	-	-
critic_score	64016	57338	10.43%	0.104318	0.104318	-	-
total_sales	64016	45094	29.56%	0.295582	0.295582	-	-
na_sales	64016	51379	19.74%	0.197404	0.197404	-	-
jp_sales	64016	57290	10.51%	0.105067	0.105067	-	-
pal_sales	64016	51192	20.03%	0.200325	0.200325	-	-
other_sales	64016	48888	23.63%	0.236316	0.236316	-	-
release_date	64016	7051	88.99%	0.889856	0.889856	-	-
last_update	64016	46137	27.93%	0.27929	0.27929	-	-

نکات)

- Currentness: در حالت کلی باید دیتاست با یک منبع مطمئن چک شود تا اطمینان حاصل کنیم که داده ها به روز هستند. اما از آن جایی که دیتاست داده شده یکتاست و نیز به هیچ گونه منبع خارجی دسترسی نداریم، فرض می کنیم دیتا ها به روز هستند.
- Consistency:

```
Consistency Ratio: 0.96
Number of Inconsistent Rows: 2775
Inconsistent Rows:
                                                 img  ... last_update
26     /games/boxart/full_5970958AmericaFrontccc.jpg  ...  2018-09-12
157    /games/boxart/full_2414906AmericaFrontccc.jpg  ...  2018-02-28
435    /games/boxart/full_4884699AmericaFrontccc.jpg  ...  2018-09-12
440    /games/boxart/full_3865712AmericaFrontccc.jpg  ...  2018-02-27
457    /games/boxart/full_9751338AmericaFrontccc.jpg  ...  2018-04-08
...                                                  ...  ...         ...
64001    /games/boxart/full_92687AmericaFrontccc.jpg  ...  2018-12-17
64002  /games/boxart/full_7615819AmericaFrontccc.png  ...  2018-12-17
64012  /games/boxart/full_8031506AmericaFrontccc.jpg  ...  2020-05-09
64013  /games/boxart/full_6553045AmericaFrontccc.jpg  ...  2020-05-09
64014    /games/boxart/full_6012940JapanFrontccc.png  ...  2019-02-24

[2775 rows x 14 columns]
```

برای بررسی ناسازگاری میان داده‌های release_date و last_update، می‌توان مقادیر دو ستون را با هم مقایسه کرده و هر ردیفی که تاریخ release_date از last_update بعدتر است را پیدا کرد. این ردیف‌ها را می‌توان به عنوان ناسازگاری‌ها مشخص کرد و میزان آنها را نیز به عنوان یک معیار برای consistency در نظر گرفت.

در اینجا، consistency ratio نسبتی است که نشان می‌دهد چه قدر از داده‌ها مطابق با قوانین یا شرایط مشخص شده هستند. به عبارت دیگر، این نسبت نشان می‌دهد چه قدر از داده‌ها به طور کلی با قوانین مورد انتظار سازگاری دارند.

اگر consistency ratio برابر با ۱ باشد، نشان می‌دهد که تمام داده‌ها با قوانین مورد انتظار سازگاری دارند. اگر کمتر از ۱ باشد، این نشان می‌دهد که یک بخشی از داده‌ها با قوانین مشخص شده سازگاری ندارند.

تابع check_consistency فقط ردیف‌هایی را که تاریخ آپدیت قبل از انتشار است را بازمی‌گرداند. این ردیف‌ها نمایانگر مواردی هستند که با قوانین یا شرایط سازگاری ندارند.

مشکلات)

- چون دیتاست یکتاست و رکوردها در شرایط مختلف، متفاوت نیستند مشکلات مربوط به multi-instance نداریم.
- مشکلات single-schema:
  - Img: بهتر است نوع آن نیز مشخص شود.
  - تعداد مقادیر غیرنالی (non-null) در هر ستون: تعداد مقادیر غیرنالی در هر ستون متفاوت است، این نشان می‌دهد که داده‌ها با ساختار یکسان در دیتاست وجود ندارند. این موضوع می‌تواند به عنوان یک نشانه از مشکلات single-schema مطرح شود.
  - مقدار غیرقانونی (مقدار تاریخ خارج از محدوده)
  - **Conflicts** (تاریخ انتشار و آخرین به روز رسانی)
- مشکلات single-instance:
  - last_update , release_date : هردو از جنس تاریخ هستند اما در تایپ داده object است
  - **Missing values**
- مشکلات **multi**_schema :
  - **تضادهای نامگذاری (مقدار متفاوت برای ژانر،...)**

رفع مشکلات)

- در خصوص رفع ایرادات گفته شده می‌توان از روش های حذف یک ستون،اضافه ستون جدید در صورت دسترسی به شما، رندسازی مقادیر و تغییر واحد یا scale یک ویژگی استفاده کرد.
- برخی رکورد ها در ستون های تاریخ انتشار و آخرین آپدیت مقادیر خارج از محدوده دارند. برای حل این مشکل می توان با بررسی جدول داده دوم، رکورد هایی که خطا دارند و به اشتباه خارج از محدوده وارد شده اند را شناسایی و تصحیح کرد. همچنین می توان رکورد دارای این نوع خطا را از نوع null در نظر گرفت و ادامه محاسبات را روی بقیه داده ها انجام داد.

با بررسی رکوردهای موجود در جدول بازی‌ها، رکوردهایی هستند که تاریخ انتشار بازی آنها بعد از تاریخ اپدیت قرار دارد. می توان با تغییر تاریخ انتشار بازی به مقدار صحیح، رکوردها را اصلاح کرد. همچنین می توان در صورت جابه جا وارد کردن داده ها، تاریخ انتشار بازی را با تاریخ آخرین اپدیت جایگزین کرد.

فاز چهارم) ترکیب دو دیتاست موجود با یکدیگر

برای رفع ناسازگاری‌ها، می‌توان روش‌های مختلفی را انتخاب کرد. یکی از راه‌های معمول، استفاده از مقادیر پیش‌فرض یا میانگین‌هاست. همچنین برای رفع ناسازگاری‌ها یک روش استفاده از یک دیتا ست خارجی می‌باشد.