

Лабораторная работа №5

Цель задания: исследовать приложенные датасеты для поиска 2 инсайтов по данным (Data Insights). Инсайт - это представление данных в виде визуализации с применением предварительной группировки данных, которое представляет интересные моменты в данных. (например, количество товара, которое было продано в прошлом сезоне).

Очень тяжело работать на стенде, ссылка на ноутбук <https://www.kaggle.com/code/saspav/recommended-game-genres>

В медианном значении: те, кто не рекомендуют игру играют в неё около 12 часов. Если игра затянула то, тот кто её рекомендовал - играл около 32 часов. В среднем это цифры 83 и 107 часов соответственно, но средние значения выглядят подозрительно из-за больших максимальных значений.

SPARK JOB FINISHED

```
%spark.pyspark

df_grp = data.groupBy('app_id', 'is_recommended').agg(
    F.mean('hours').alias('hours_mean'),
    F.expr('percentile_approx(hours, 0.5)').alias('hours_median'),
    F.countDistinct('user_id').alias('users')
)

df_grp = df_grp.fillna(0)

data.groupBy('is_recommended').agg(
    F.expr('percentile_approx(hours, 0.5)').alias('median_hours'),
    F.mean('hours').alias('mean_hours'),
    F.max('hours').alias('max_hours')
).show()
```

is_recommended	median_hours	mean_hours	max_hours
false	12	83.19694173228763	998
true	32	106.7596313227745	998

Took 14 sec. Last updated by pavlov at August 06 2023, 6:37:09 PM.

```
%spark.pyspark
```

```
# Только одна игра, которая входит в TOP-10 по количеству играющих и затраченных часов
set(title_users) & set(title_hours)

{'Counter-Strike: Global Offensive'}
```

Took 1 min 5 sec. Last updated by pavlov at August 06 2023, 8:55:58 PM.

```
%spark.pyspark
```

```
# TOP-10 игр, за которыми было затрачено больше всего времени в среднем
top = df.sort(F.desc("hours_mean")).limit(10)
top.show(10)
```

app_id	is_recommended	hours_mean	hours_median	users
570	true	444.2949113978709	430	120426
730	true	429.0136694202447	397	186182
730	false	423.1161377245509	391	33400
1283970	true	416.0554371002132	383	469
236850	true	409.83787078423404	998	36915
401090	true	404.59090909090907	299	22
1011510	true	401.30021141649047	344	473
39210	true	398.24465990325757	348	46722
944770	true	394.9577464788732	323	71
1097960	true	391.7391304347826	329	46

Took 4 min 49 sec. Last updated by pavlov at August 06 2023, 8:54:14 PM.

```
%spark.pyspark
```

```
# TOP-10 игр, в которых играло больше всего пользователей
top = df.sort(F.desc("users")).limit(10)
top.show(10)
```

app_id	is_recommended	hours_mean	hours_median	users
440	true	324.299959631322	250	294782
252490	true	361.12425507082355	297	226196
431960	true	102.60637631501459	37	186785
730	true	429.0136694202447	397	186182
374320	true	174.32732812970784	125	175903
227300	true	180.86381976350958	100	171508
550	true	137.43974943517833	51	171293
444090	true	137.90052567022954	57	171210
1091500	true	112.98926637766932	85	168629
359550	true	360.95677597176785	301	155284

Took 4 min 9 sec. Last updated by pavlov at August 06 2023, 8:32:51 PM.

```
%spark.pyspark
```

```
top_games = [row.app_id for row in top.select('app_id').collect()]
app_games = games.filter(F.col("app_id").isin(top_games)).select('app_id', 'title')
app_games.show()
```

app_id	title
1011510	Wizard And Minion...
944770	sheepChat
1283970	YoloMouse
1097960	ClickRaid2
401090	MODO indie
570	Dota 2
730	Counter-Strike: G...
39210	FINAL FANTASY XIV...
236850	Europa Universali...

Took 2 min 31 sec. Last updated by pavlov at August 06 2023, 8:54:53 PM.

```
%spark.pyspark
```

```
top_games = [row.app_id for row in top.select('app_id').collect()]
app_games = games.filter(F.col("app_id").isin(top_games)).select('app_id', 'title')
app_games.show()
```

app_id	title
440	Team Fortress 2
550	Left 4 Dead 2
730	Counter-Strike: G...
227300	Euro Truck Simula...
252490	Rust
359550	Tom Clancy's Rain...
374320	DARK SOULS™ III
431960	Wallpaper Engine
444090	Paladins®
1091500	Cyberpunk 2077

Took 27 sec. Last updated by pavlov at August 06 2023, 8:43:34 PM.