게임 유저 데이터 분석

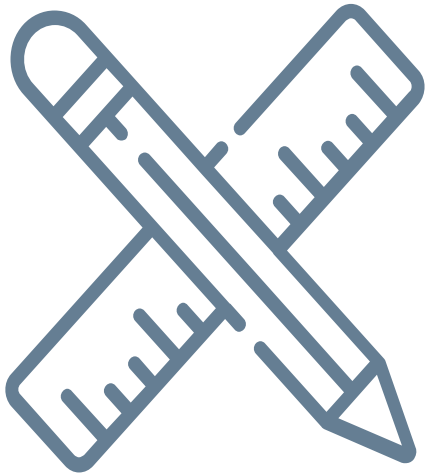
탐색적 데이터 분석(EDA)

목표 및 목차

- 1주일간의 게임 데이터를 분석해 개선할 사항을 찾고 그에 따른 개선 방안을 마련

1. 데이터 전처리

colab으로 주어진 csv파일을 불러와
데이터 전처리 실행



컬럼명 수정

```
df = df.rename(columns={'countents1': 'contents1', 'countents2': 'contents2'})
df
```

contents1	contents2	contents3	contents4	contents5

- 이상이 있던 컬럼명을 더 수월한 작업을 위해 수정

결측값 제거

	0
userID	10
groupID	9
daycnt	0
playtime	0
payamount	0
countents1	3
countents2	13
contents3	3
contents4	3
contents5	3
dailyquest	3
weeklyquest	3
marketbuy	3
marketsell	3

- 각 컬럼에 대한 결측값의 개수를 확인한 후 결측값이 가장 많은 ‘countents’ 열에 대한 결측값이 있는 행을 출력

```
df[df['contents2'].isna()]
```

	userID	groupID	daycnt	playtime	payamount	contents1	contents2	contents3	contents4	contents5	dailyquest	weeklyquest	marketbuy	marketsell
5923	637738	5.0	2	0.883611	0.000000	0.0	NaN	0.0	0.0	3206.0	0.0	0.0	0.0	0.0
5924	FF6252	3.0	7	13.181111	0.000000	17.0	NaN	0.0	2.0	9717.0	33.0	5.0	0.0	11.0
5925	D17C11	6.0	7	46.764722	0.000000	0.0	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5926	615E2B	3.0	7	15.274444	0.000000	33.0	NaN	0.0	4.0	12960.0	9.0	0.0	4.0	14.0
5927	1B07B7	3.0	7	106.808333	0.000000	14.0	NaN	12.0	5.0	23434.0	31.0	6.0	21.0	4.0
5928	40344C	2.0	1	0.070556	0.000000	0.0	NaN	0.0	0.0	0.0	0.0	0.0	1.0	0.0
5929	B83F5E	6.0	5	58.939167	0.000000	0.0	NaN	0.0	0.0	9519.0	0.0	0.0	2.0	0.0
5930	25E2C9	3.0	7	73.785000	4.201681	17.0	NaN	0.0	7.0	32779.0	49.0	1.0	18.0	11.0
5931	AFFE47	5.0	2	0.531389	0.000000	0.0	NaN	0.0	0.0	0.0	1.0	0.0	1.0	3.0
5932	7D98B8	9.0	7	49.079444	37.957219	11.0	NaN	0.0	4.0	18947.0	5.0	0.0	251.0	25.0
6216	B23ECC	2.0	1	2.019722	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6217	A6918B	5.0	2	0.069167	114.621849	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6262	509A54	2.0	1	0.009167	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

결측값 제거

```
df[df['contents2'].isna()]
```

	userID	groupID	daycnt	playtime	payamount	contents1	contents2	contents3	contents4	contents5	dailyquest	weeklyquest	marketbuy	marketsell
5923	637738	5.0	2	0.883611	0.000000	0.0	NaN	0.0	0.0	3206.0	0.0	0.0	0.0	0.0
5924	FF6252	3.0	7	13.181111	0.000000	17.0	NaN	0.0	2.0	9717.0	33.0	5.0	0.0	11.0
5925	D17C11	6.0	7	46.764722	0.000000	0.0	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5926	615E2B	3.0	7	15.274444	0.000000	33.0	NaN	0.0	4.0	12960.0	9.0	0.0	4.0	14.0
5927	1B07B7	3.0	7	106.808333	0.000000	14.0	NaN	12.0	5.0	23434.0	31.0	6.0	21.0	4.0
5928	40344C	2.0	1	0.070556	0.000000	0.0	NaN	0.0	0.0	0.0	0.0	0.0	1.0	0.0
5929	B83F5E	6.0	5	58.939167	0.000000	0.0	NaN	0.0	0.0	9519.0	0.0	0.0	2.0	0.0
5930	25E2C9	3.0	7	73.785000	4.201681	17.0	NaN	0.0	7.0	32779.0	49.0	1.0	18.0	11.0
5931	AFFE47	5.0	2	0.531389	0.000000	0.0	NaN	0.0	0.0	0.0	1.0	0.0	1.0	3.0
5932	7D98B8	9.0	7	49.079444	37.957219	11.0	NaN	0.0	4.0	18947.0	5.0	0.0	251.0	25.0
6216	B23ECC	2.0	1	2.019722	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6217	A6918B	5.0	2	0.069167	114.621849	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6262	509A54	2.0	1	0.009167	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
# contents1~marketsell 모두 결측값인 행 제거
df = df.dropna(subset=['contents3'])
df
```

- countents1부터 marketsell 모두 결측값이 있는 데이터는 사용하기 어렵다고 판단후 행 제거

결측값 제거

```
mean_value = df['contents2'].mean()
mean_value
#contents의 값들 모두 정수임으로 반올림 후 2로 계산
```

```
1.7192645883293365
```

```
df['contents2'].fillna(2, inplace=True)
```

- 남은 10개의 ‘countents2’열의 결측값은 평균 값으로 대체

	0
userID	10
groupID	0
daycnt	0
playtime	0
payamount	0
contents1	0
contents2	0
contents3	0
contents4	0
contents5	0
dailyquest	0
weeklyquest	0
marketbuy	0
marketsell	0

- ‘groupID’의 결측값이 있는 행들은 모두 제거
- ‘userID’의 경우 group별로 데이터 EDA를 진행하는데 문제가 없기 때문에 전처리 하지 않음

결측값 제거

```
mean_value = df['contents2'].mean()
mean_value
#contents의 값들 모두 정수임으로 반올림 후 2로 계산
```

```
1.7192645883293365
```

```
df['contents2'].fillna(2, inplace=True)
```

	0
userID	10
groupID	0
daycnt	0
playtime	0
payamount	0
contents1	0
contents2	0
contents3	0
contents4	0
contents5	0
dailyquest	0
weeklyquest	0
marketbuy	0
marketsell	0

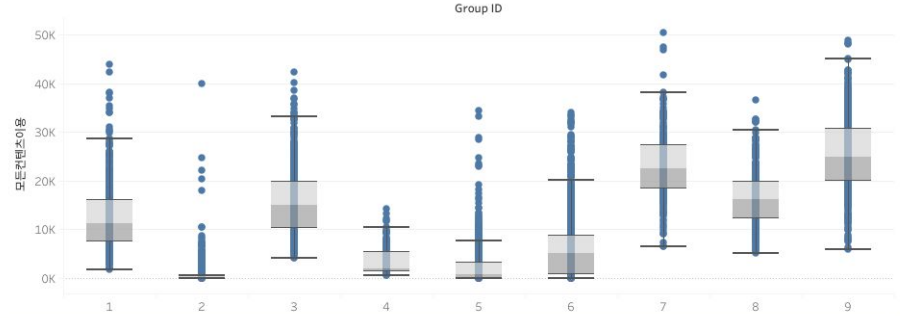
- ‘groupID’의 결측값이 있는 행들은 모두 제거
- ‘userID’의 경우 데이터 EDA를 진행하는데 문제가 없기 때문에 전처리 하지 않음

2. 데이터 시각화

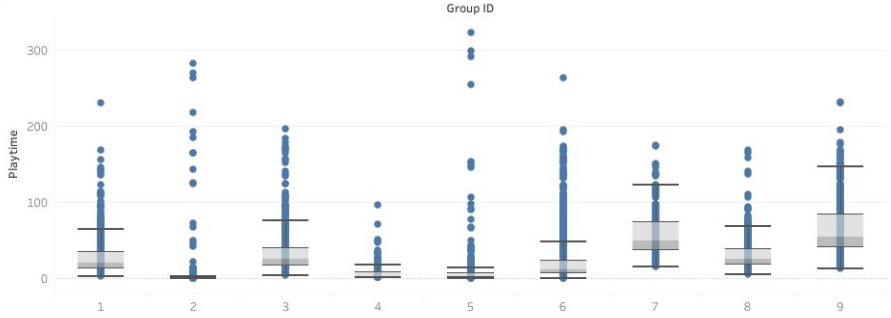
전처리한 csv파일을 Tableau와 colab(Python)을 활용하여 데이터 시각화 및 분석

이상치 분석

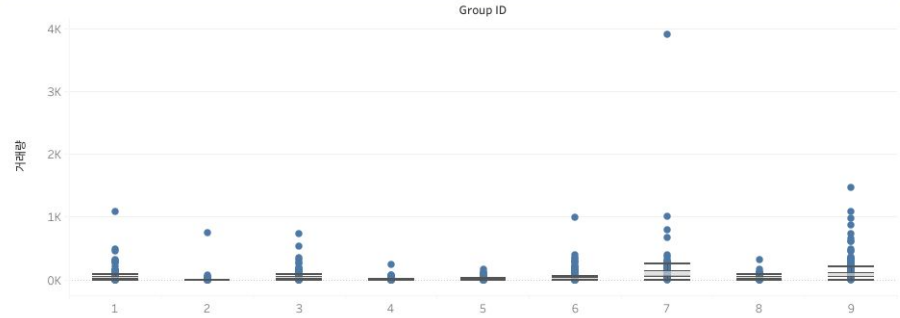
컨텐츠이용량 분포도



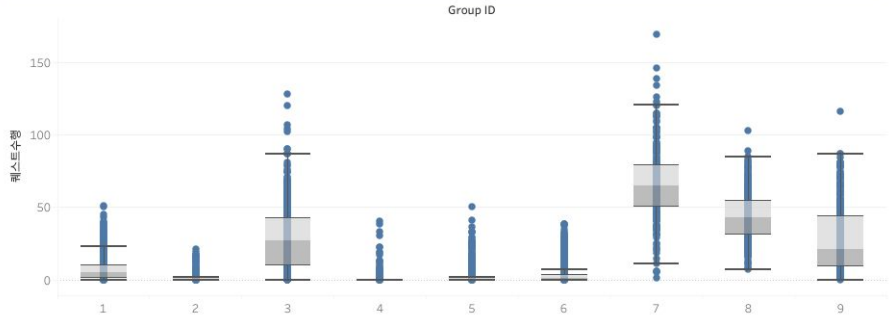
플레이시간 분포도



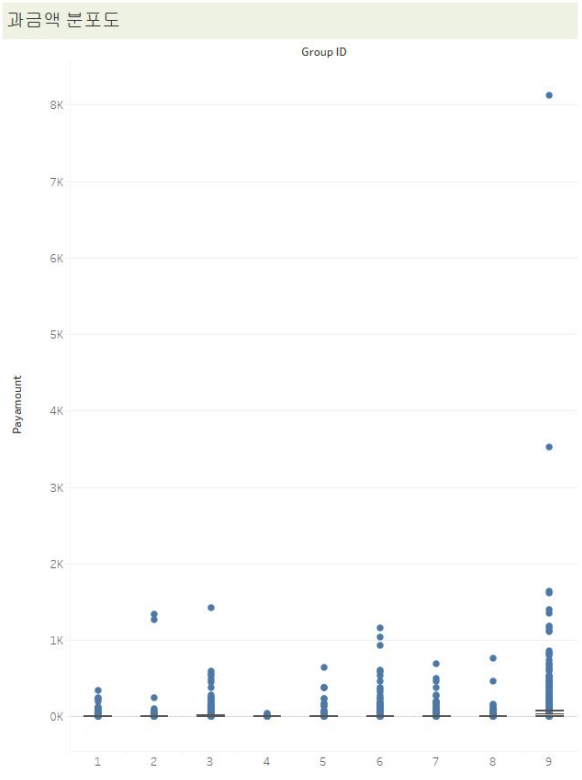
거래량 분포도



퀘스트수행량 분포도

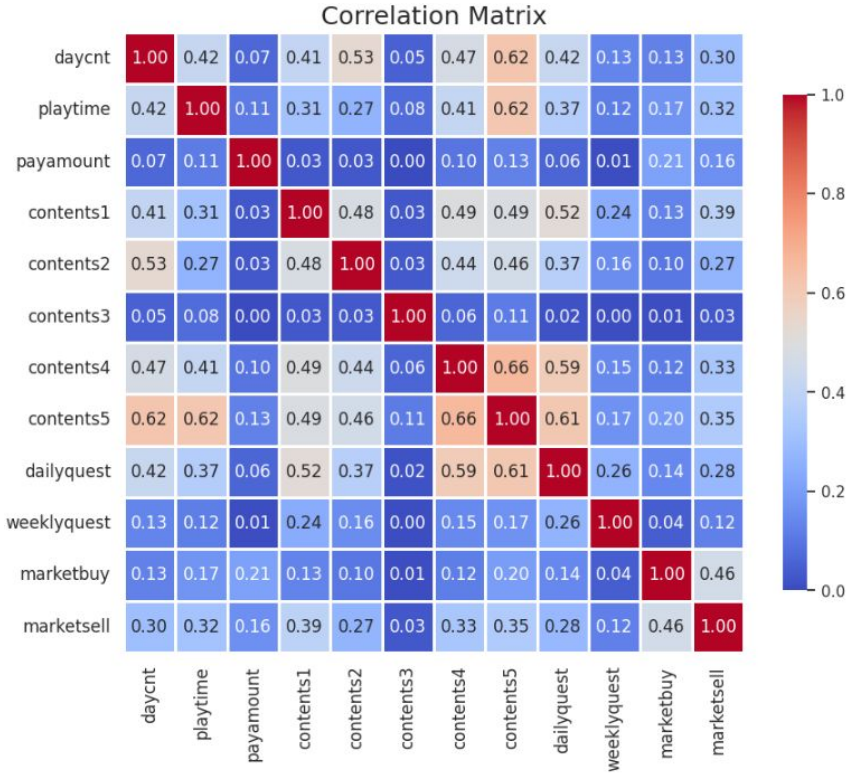


이상치 분석



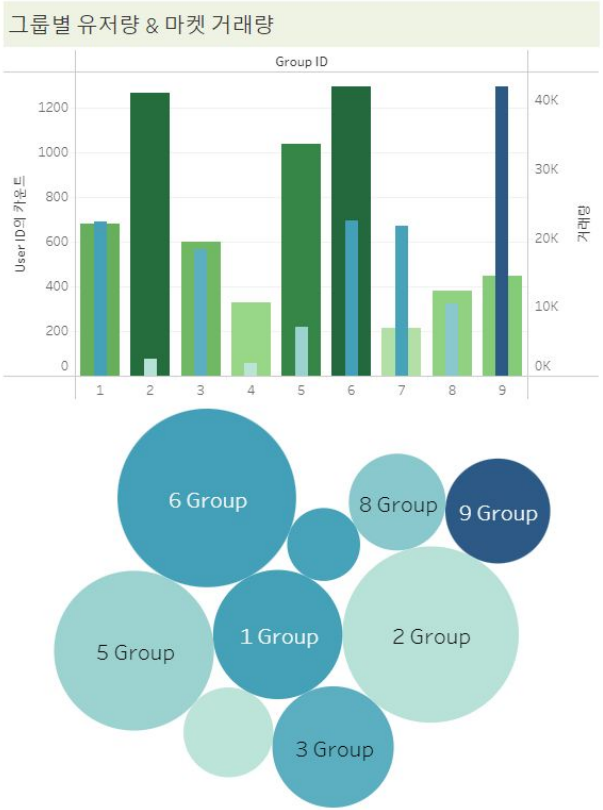
- 다음 이상치 분석 결과를 박스 플롯 형태로 보았을 때 많은 데이터가 이상치로 이루어져 있다는 것을 알 수 있다.
- 하지만 데이터의 특성상 게임을 플레이하는 개인의 차가 크므로 이상치가 많이 발생할 수 밖에 없으며 각 컬럼의 단위를 알 수 없어 섣불리 이를 처리하는 것은 바람직 하지 않다고 생각된다.

상관관계분석



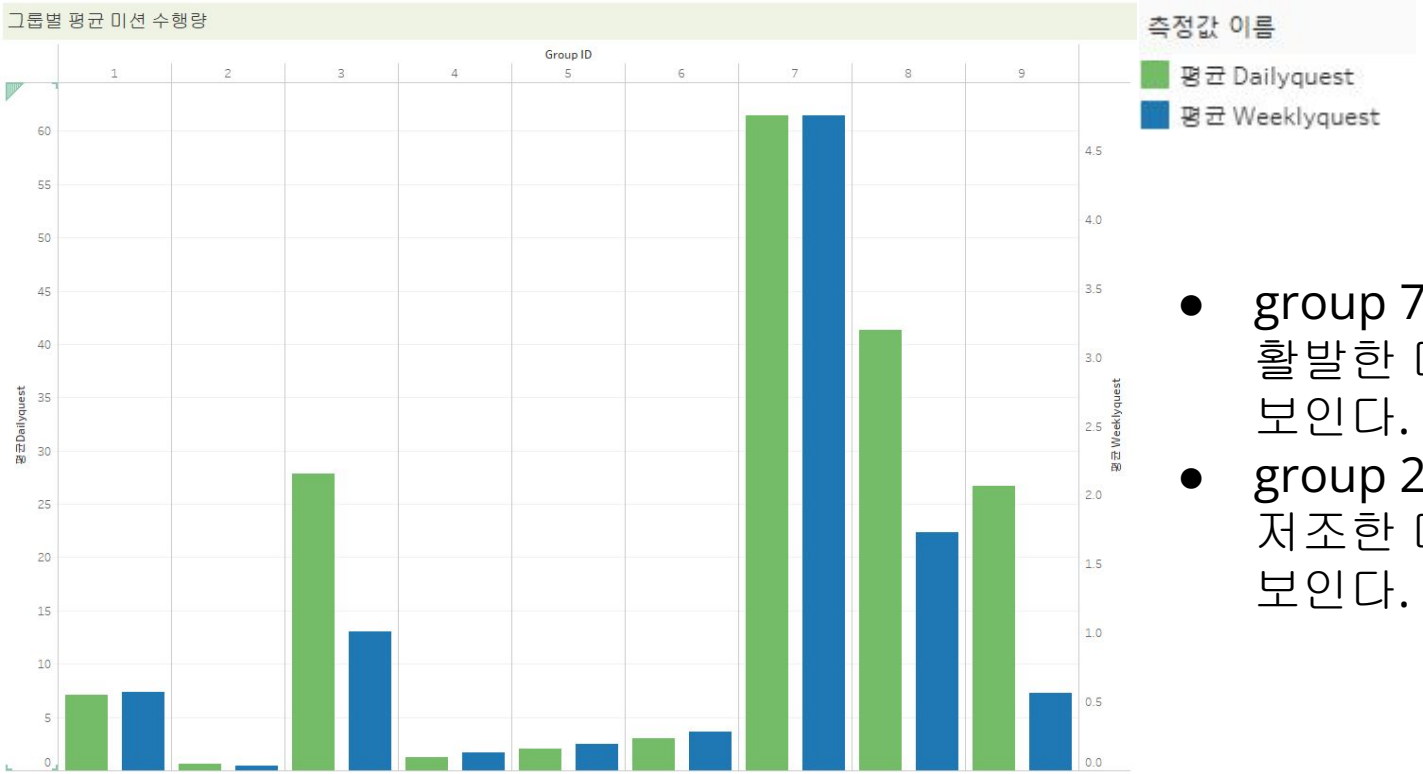
- 각각의 상관계수를 시각화한 결과
- 위처럼 강한 상관이 있는 관계는 없으며 contents5와 daycnt, playtime, contents4, dailyquest 만이 상관이 있다고 보여진다.
- 약한 상관이 있는 관계로는 0.4이상 0.6 이하의 상관계수가 있는 관계들이 있다.

그룹별 유저량 & 마켓 거래량



- 위 그래프로 그룹의 활성화 정도를 판단할 수 있다.
- group 2, 5, 6은 많은 유저를 보유하고 있다.
- group 7, 9는 활발한 마켓 거래량을 보여준다.
- group 2, 4, 5는 유저량에 비해 저조한 마켓 거래량이 활성화 되어 있지 않다.

그룹별 1인당 일간 & 주간 미션 수행량

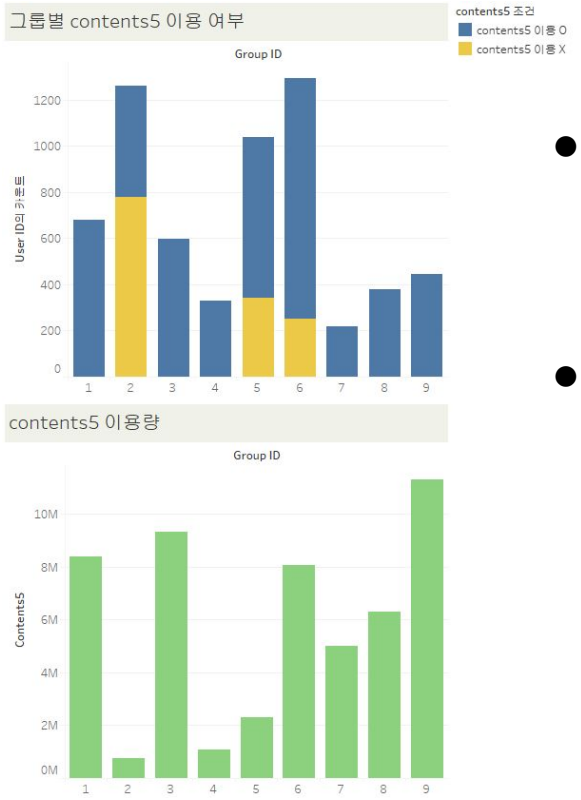


- group 7에서 가장 활발한 미션 수행량을 보인다.
- group 2, 4, 5, 6에서 저조한 미션 수행량을 보인다.

그룹별 컨텐츠 이용량

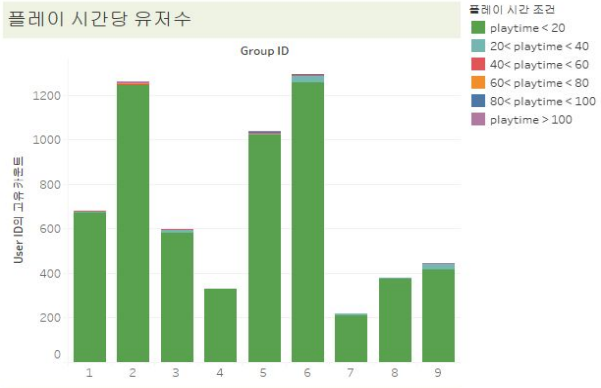
컨텐츠 이용률 수	
Contents1	147,971
Contents2	10,774
Contents3	1,390
Contents4	10,703
Contents5	52,468,272

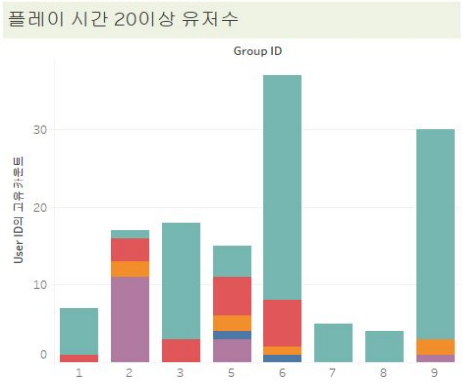
- contents1~4는 전체적인 이용률이 저조해 의미있는 분석이 나오기 힘들다.



- group 2, 5, 6을 제외한 모든 그룹의 유저는 contents5을 이용한다.
- group 2에서는 과반 이상이 contents5을 이용하지 않는다.

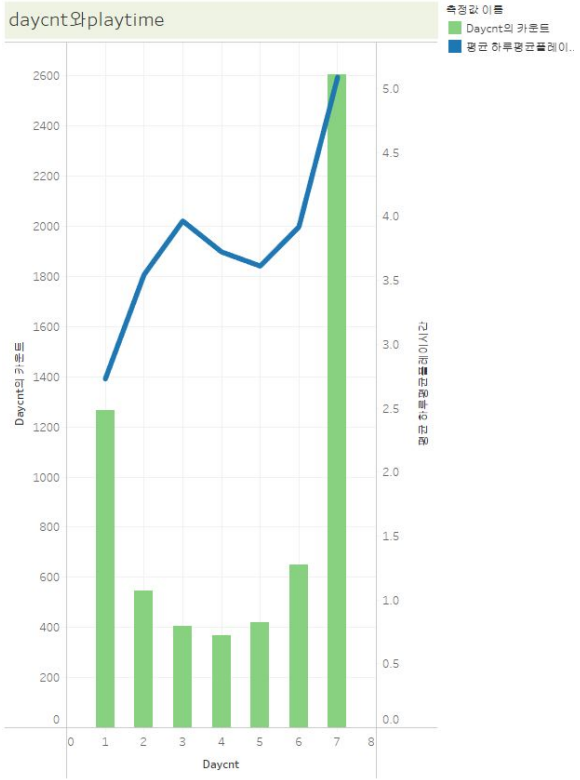
그룹별 플레이 시간당 유저 수





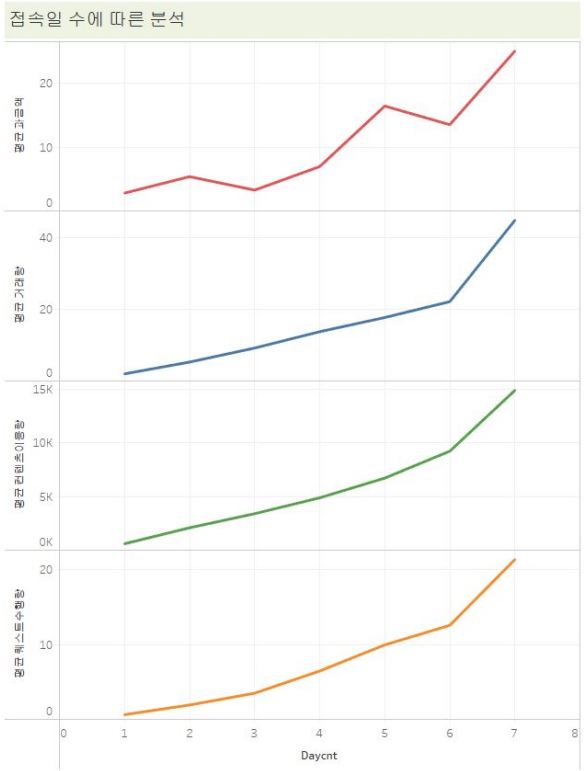
- 모든 그룹 유저 수의 대부분이 playtime이 20보다 적다.
- group 4는 모든 유저가 20보다 적은 playtime을 가지고 있다.
- group 2와 5에서 많은 playtime을 가진 유저들이 많다.

접속일 수와 플레이 시간



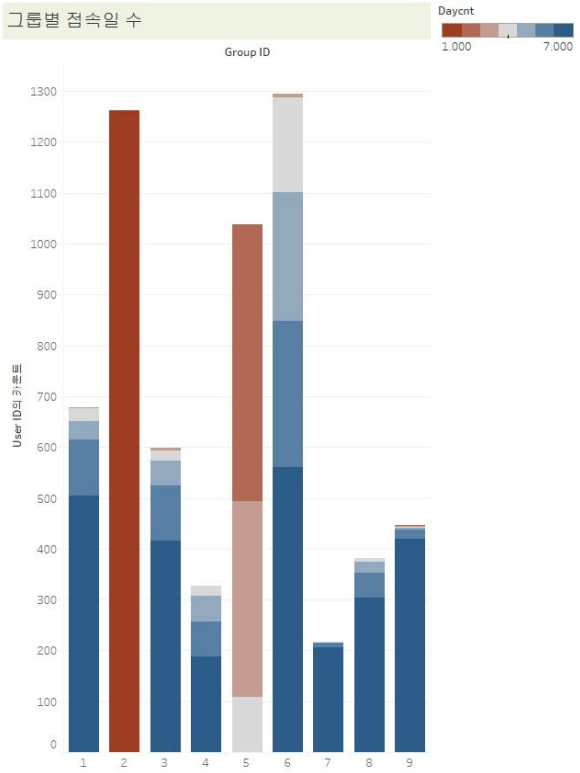
- 접속일 수가 많은 유저일수록 하루 평균 플레이 시간이 늘어나는 경향을 보인다.
- 주 7일 모두 플레이하는 유저의 비율이 가장 높다 -> 접속일 수는 게임에 대한 유저들의 충성도를 나타낸다고 예상.

접속일 수와 플레이 시간



- 이전의 예상대로 접속일 수에 따라 평균 과금액, 거래량, 컨텐츠 이용량, 퀘스트 수행량이 비례하는 경향을 보인다.
- 7일을 모두 접속한 유저들의 경우 모든 분야에서 다른 유저들에 비해 월등한 이용률을 보인다.
- 접속일 수는 유저들의 충성도를 나타낸다.

그룹별 유저의 접속일 수



- group 2와 5는 대부분의 유저가 접속일 수가 저조하다.
- 나머지 group들은 많은 접속일 수를 기록한 유저가 많다.

3. 결론 도출

데이터들을 분석한 결과를 바탕으로
개선 사항과 개선 방법을 제시

결론 도출

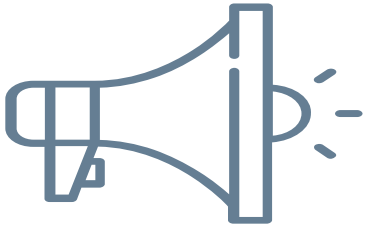
- 유저들의 플레이 시간을 늘리는 것 보다 접속일 수를 늘리는 것이 게임의 활성화에 더욱 도움이 된다.

➔ 이벤트를 진행할 때 플레이 시간에 따라 보상하는 방안 보다 출석률에 따른 보상을 하는 방안이 더욱 효과적일 것이다.

결론 도출

- contents 5를 제외한 나머지 콘텐츠에 대한 모든 그룹 유저들의 이용률이 매우 저조하다.

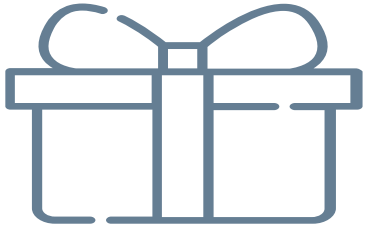
➔ 새로운 콘텐츠를 출시하여 새로운 유저 혹은 라이트 유저들의 관심을 끌 수 있다.



결론 도출

- 특정 group을 제외하고 미션수행률이 현저히 낮다

➔ 미션에 대한 보상을 늘려 미션 수행률을 높일 수 있을 것이다.
➔ 하지만 보상이 너무 높아질 경우 유저들의 과금액이 줄어들 수 있다.



추가로 필요한 데이터 및 정보

- 그룹의 구분 조건, 콘텐츠의 종류, 미션의 종류
  ➔ 더욱 세밀하고 정확한 데이터 분석이 가능해진다.

- 연속형 변수들에 대한 단위 (플레이시간, 과금액, 콘텐츠 이용량, 미션 수행량, 구매량, 판매량)
  ➔ 데이터에 오류가 있는지 확인 할 수 있다.

- 유저의 게임 가입일
  ➔ 유저가 얼마나 해당 게임을 이용했는지 알 수 있으며 신규 유저와 올드 유저들을 구분하여 데이터를 분석할 수 있다.

감사합니다


김호원


ghdnjs0921@naver.com


010-3310-6524