네 안녕하세요
13기 I반 오봉석입니다.
섹션1 프로젝트 발표를 시작하겠습니다

먼저 데이터를 불러옵니다.

Null 값과 데이터 타입을 확인합니다.
결측치가 10프로 미만이면 삭제나 대치해도 됩니다
Name feature은 unique값이 너무 많아서 쓸모 없다고 판단했습니다.
그래서 필요없는 결측치와 feature2개를 처리했습니다.
잘 되었는지 확인까지 해줬구요

Sales feature들을 하나씩 들여다보니 일단 데이터 타입은 string이였고 K,와 M이 섞여있었습니다.
그래서 가장 기본인 밀리언으로 값들을 대치해줬습니다.

apply를 한번에 하고 싶었는데 반복 메서드를 또 반복해주면 오류가 뜨길래 하나씩 적용했습니다.
사용자 함수를 적용해 K와 M이 있던 데이터는 float으로 만져줬고 나머지 데이터에 대해서 pd.to_numeric으로 타입을 변경해줬습니다.

질문1에 대한 과정을 진행하기 위해 판매량과 장르만 뽑아서 데이터프레임을 만들었습니다.
장르를 그룹으로 각 지역별 판매량 합계를 나타냈습니다.
시각화를 편하게 하기 위해 타이디 데이터로 만들었는데 생각보다 시간이 오래 걸려서 필요한 데이터만 추출했습니다.

각 지역마다 판매량이 많도록 데이터를 정렬하고, 그때 탑5 장르를 데이터프레임으로 뽑아냈습니다.
그 후 JP에서 롤플레잉 제외시 대부분 지역에서 상위 랭크된 게임 장르는 같다고 나왔습니다.

질문 2에 대해서 진행하겠습니다.
사실 질문에 대한 고민을 꽤 오래 했는데 단순히 연관성 유무만 도출하면 되는지, 아니면 어떤 연관성이 있는지도 알아내야 하는지 갈등했씁니다.

연도도 숫자가 4자리인 경우와 2자리인 경우가 있어서
처음에 숫자 2자리인 것들을 1900과 2000을 붙여준 형태로 바꿔보려 했으나 계속 에러가 떠서 실패했씁니다.
그래서 두자리인 경우는 0.6프로 정도길래 그냥 결측치로 보고 제거했습니다.

연도가 continuous이므로 categorical로 바꿔주는 함수를 적용해줬고
카테고리마다 데이터 개수도 확인해줬습니다.

그 후 전 지역을 합해서 토탈 판매량 Feature을 만들어줬습니다.
게임회사와 플랫폼에 대해서도 통계치를 요약해봤습니다.
연도별 트렌드를 알기 위해 플랫폼, 퍼블리셔, 장르에 대해 pd.crosstab 하고 차이스퀘어

투샘플 테스트로 연관성이 잇는지 진행했습니다.

P value가 모두 0이거나 근처여서 3개 피쳐가 연도와 연관성이 있다고 합니다.
연도에 대해 그룹바이하면 1980, 1990, 2000, 2010년대로 나뉘는데, 그룹당 5프로에 해당하는 개수를 n에 지정해줬습니다.
그 후 그룹마다, 판매량이 많은 기준으로 개수 n에 해당하는 만큼 데이터프레임을 할당해줬구요

일단 트렌드라는 기준은 판매량이므로, 이 필터링된 데이터를 관측하면 트렌드가 보일거라 판단했씁니다.

그래서 플랫폼에 대해 연도마다 상위 5개에 대해 비율로 뽑아냈습니다.
적힌 설명 진행

사실 이것도 시각화를 하고 싶었는데 아직 시각화 라이브러리가 미숙한지 눈으로 확인되는 유동성만 봤습니다.

퍼블리셔에 대해서도 진행했구요
설명 진행

장르에 대해서도 진행

이제 3번 문제인데요,

아까 만들었던 총 판매량에 대해 정렬을 했습니다.
이 데이터를 다 쓰기엔 또 좋은 데이터가 아니어서
판매량이 1보다 큰 것으로 필터링을 걸었습니다.
2009개 정도가 나오네요

데이터에 대해 플랫폼으로 그룹바이 하고 통계치를 확인했습니다. 저희가 궁금한건 총 판매량과 어떤 상관성이 있는지가 궁금한데요.

통계치에서 count(개수)와 mean(평균)을 곱해서
플랫폼의 unique값마다 총합계 판매량을 구하려고 했습니다.
sum이라는 feature에 할당하고 오름차순으로 정렬해줬구요
그 후 상위 랭크 10까지를 시각화했습니다.

플랫폼에서는 전 시대에 아우러 PS2가 가장 많이 팔렸고, X360, PS3이 뒤따르고 있습니다

장르는 액션이 나름 격차를 갖고 1위를 하고 스포츠와 슈터가 2,3위를 다투고 있습니다.

음 게임회사에서 닌텐도가 2010년대 판매량에서 2위로 내려왔는데 옛날 판매량이 많다보니 압도적 1위가 되었습니다.

제가 분석한건 여기까지입니다. 제가 스스로 제기한 문제가 없어서 아쉬움이 남습니다. 다음엔 더 잘해야겠다 다짐합니다. 감사합니다