AI 18기 최민석, Section1 project 발표를 시작하겠습니다.

목차입니다. ‘Data preprocessing & EDA’, ‘지역에 따라 선호하는 게임 장르에 대한 분석’, ‘연도별 게임 트렌드에 대한 분석’, ‘다음 분기에 설계해야 할 게임’ 순으로 설명하겠습니다.

먼저 Data preprocessing & EDA 입니다.

정제하지 않은 데이터프레임 입니다. 이 데이터프레임은 16898개의 행과 10개의 열로 이루어져 있습니다. 보시는 것과 같이 Unnamed : 0 라는 의미 없는 열이 있으며, ‘Year’ 열의 경우는 실수형으로 표시 되어있고, ‘Sales’ 열에는 480K, 0.33M 과 같이 숫자가 아닌 문자들도 포함되어 있는 것을 볼 수 있습니다. 따라서 데이터를 분석하기 전에 데이터프레임을 정제하는 Data preprocessing & EDA 과정을 진행해야합니다.

우선 데이터프레임의 결측치를 확인합니다. 확인 결과 ‘Year’ 열에서 271개, ‘Genre’ 열에서 50개, ‘Publisher’ 열에서 58개의 결측치가 확인되었습니다. 데이터프레임의 전체 데이터 수에 비해 결측치가 존재하는 데이터의 수는 작으므로 제거해도 데이터분석에 큰 영향을 끼치지 않을 것 같다는 판단을 하여 결측치가 존재하는 행들은 모두 삭제하기로 결정했습니다. 또한 ‘Unnamed : 0’ 열은 의미가 없는 열이기 때문에 열 전체를 제거합니다.

‘Year’ 열의 데이터들은 실수형으로 이루어져 있기 때문에 다음과 같은 코드를 작성해 이를 정수형으로 고쳐줍니다. Sales 열들의 기본단위는 조사결과 M(백만) 입니다. 이 열들에 존재하는 K 와 M 이라는 문자는 각각 ‘천’ 과 ‘백만’ 을 의미한다고 생각할 수 있기 때문에 K 가 붙어있는 데이터의 경우 기본단위인 백만으로 맞춰주기 위해 K를 제거하고 0.001을 곱해줍니다. M 이 붙어있는 데이터의 경우 M을 제거합니다. 마지막으로 데이터타입을 실수형으로 고쳐줍니다.

이 과정들을 거치게 되면 다음과 같이 데이터프레임의 데이터들이 깔끔하게 표시되는 것을 확인할 수 있습니다. 이제 각각의 열들을 살펴봄으로써 데이터에 이상이 없는지 확인합니다.

‘Name’ 열의 이상치를 확인합니다. ‘Name’, ‘Platform’, ‘Year’, ‘Genre’, ‘Publisher’ 열이 모두 중복인 값을 찾고 병합합니다. 다음과 같이 병합은 ‘Name’, ‘Platform’, ‘Year’, ‘Genre’, ‘Publisher’ 은 하나로 합치고 ‘Sales’ 열들은 중복인 값들의 데이터들을 더해주는 방식으로 진행했습니다. 나머지 데이터들에서는 'Name' 이 중복되어도 'Platform', 'Year', 'Genre' 에서 다른 값들을 가지고 있으므로 더이상 중복된값들은 제거하지 않기로 결정했습니다.

‘Platform’ 열의 이상치를 확인합니다. Value\_counts() 매서드를 활용해 ‘Platform’ 열의 데이터를 분석했습니다. 개수가 적은 3DO, TG16, GG, PCFX 도 찾아본 결과 실제 존재하는 게임 플랫폼이므로 'Platform' 열에서는 이상치가 없다고 판단했습니다.

‘Year’ 열의 이상치를 확인합니다. ‘Year’ 열의 값들을 내림차순, 오름차순으로 정렬해본 결과 년도의 기본적인 형태인 19xx, 20xx 가 아닌 숫자들도 포함되어 있는 것을 확인할 수 있습니다.

0~20 까지의 값들은 2000년~2020년이라는 값으로 추측, 80~99 까지의 값들은 1980년~1999년이라는 값으로 추측하여 ‘Year’ 열의 값을 1900이상, 0~20, 80~99 로 분류해본 결과 전체 데이터 16240개에서 1900이상에 16140개, 0~20에 86개, 80~99에 14개가 있음을 확인할 수 있습니다. 따라서 0~20 사이의 숫자는 2000년대, 80~99 사이의 숫자는 1900년대로 바꾸는 것이 타당합니다.

‘Genre’ 열의 이상치를 확인합니다. Value\_counts() 매서드를 활용해 확인해본 결과 이상치가 없다고 판단했습니다.

‘Publisher’ 열의 이상치를 확인합니다. ‘Publisher’ 열에서 Unknown인 값들을 추출했습니다. 'Publisher' 열에서 Unknown인 열이 99개가 존재하는데 나머지 값(Name, Platform, Year 등등)들은 유효하다고 판단하여 제거하지 않기로 결정했습니다.

Decade 라는 새로운 열을 추가해 Data Preprocessing & EDA 가 끝난 최종 데이터프레임을 완성했습니다.

다음은 지역에 따라 선호하는 게임 장르에 대한 분석입니다.

지역에 따라 선호하는 게임 장르에 대해 분석하기 위해 필요한 'Genre', 'NA\_Sales', 'EU\_Sales', 'JP\_Sales', 'Other\_Sales' 열만 뽑아 새로운 데이터프레임을 만들었습니다. ‘판매량이 높을수록 선호도가 높다’ 는 기준을 가지고 지역에 따라 장르별 게임 판매량을 확인했습니다. 또한 지역별 게임판매량의 총합을 구해봤습니다. 이를 통해 시장의 규모는 북미지역, 유럽지역, 일본지역, 기타지역 순으로 크다는 것을 파악할 수 있습니다.

시각화를 통해 지역에 따라 어떤 장르의 게임이 잘 팔리는지 쉽게 확인할 수 있습니다. 일본지역을 제외하고 나머지 지역에서 Action, Sports, Shooter 순으로 판매량이 높다는 것을 확인할 수 있습니다.

지역에 따라 선호하는 게임 장르를 수치화해서 표현했습니다. 일본지역을 제외하고는 Action, Sports, Shooter 순으로 판매량이 높은 것을 확인할 수 있고, 그 중에서도 북미지역의 경우 판매량의 개수 자체가 다른 지역들에 비해 월등히 높았습니다. 일본지역에서는 Role-Playing, Action, Sports 순으로 판매량이 높습니다.

다음은 연도별 게임 트렌드에 대한 분석입니다.

연도별 게임의 트렌드를 파악하기위해 필요한 'Genre', 'NA\_Sales', 'EU\_Sales', 'JP\_Sales', 'Other\_Sales', 'Decade' 열만 뽑아 데이터프레임을 만들었습니다. 그리고 ‘Genre’ 와 ‘Decade’ 를 groupby() 매서드를 활용해 묶어 새로운 데이터프레임을 만든 뒤, 'NA\_Sales', 'EU\_Sales', 'JP\_Sales', 'Other\_Sales' 합을 나타내는 ‘Total\_Sales’ 라는 새로운 열을 만들어줍니다. 이 데이터프레임을 가지고 Decade 에 따라 총 판매량을 확인할 수 있는 막대그래프를 그렸습니다. 이 그래프를 통해 1980년대부터 게임의 판매량이 증가하다가 2000년대에 최고점을 찍고 다시 내려가고 있는 추세임을 알 수 있습니다.

다음은 Decade별 선호하는 게임 장르 비율을 확인하기 위해 파이그래프를 그렸습니다. 그래프에서 빨간색(Platform) 장르는 비율이 점점 줄어들고, 분홍색(Puzzle) 장르는 1990년대부터 사라졌으며, 노란색(Action), 주황색(Shooter), 초록색(Sports) 장르는 비율이 점점 증가하는 것을 확인할 수 있습니다.

Decade에 따라 장르별 게임 판매량 변화를 알아보기 위해 각 장르별 꺾은선 그래프를 그려 확인해봤습니다. 총 12개의 장르를 그래프 변화를 기준으로 성장 가능성이 있는 장르와 성장 가능성이 없는 장르로 나누어 표현해봤습니다. 이 중 가장 중요하게 봐야하는 장르는 Shooter와 Action입니다. 먼저 Shooter 장르의 경우 2010년대는 2000년대에 비해 총 게임 판매량이 줄었음에도 그래프가 우상향 그래프를 나타내고 있습니다. Action 장르의 경우 게임 판매량이 다른 장르들에 비해 월등히 높습니다.

마지막으로 다음 분기에 설계해야 할 게임에 대해 설명하겠습니다.

지역별, 연도별로 게임 장르에 대해 분석한 결과를 살펴보겠습니다. 먼저 지역별 총 게임 판매량과 Decade에 따른 총 게임 판매량을 보겠습니다. 이를 통해 북미시장의 규모가 가장 큰 것을 알 수 있고, 2000년대까지 게임 판매량이 증가하다가 그 이후부터 줄어들고 있는 것을 알 수 있습니다.

다음으로 2010년대에 선호하는 게임 장르 비율을 나타내는 파이그래프와 Action 장르와 Shooter 장르의 꺾은선 그래프를 확인해보겠습니다. 2010년대에 가장 인기가 있던 장르는 Action 장르와 Shooter 장르였습니다. 또한 두 장르는 총 판매량을 고려했을 때 계속해서 인기가 많아지고 있음을 확인할 수 있습니다.

따라서 다음 분기인 2023년 4월에 설계해야 할 게임은 Action 장르 또는 Shooter 장르입니다. 이 후 시장의 규모가 가장 큰 북미시장을 공략하는 것이 가장 좋습니다. 하지만 2000년대 이후 게임 판매량이 감소하고 있는 것을 감안하여 과감한 투자보다는 신중한 투자가 중요합니다.