안녕하세요. 게임 데이터 분석을 통한, 다음에 설계할 게임 찾기 프로젝트를 발표할 장혜수입니다.

발표는 다음과 같은 순서로 진행하겠습니다.

게임 회사 데이터 팀의 일원으로서, 기존의 게임 데이터를 분석하여 다음 분기에 어떤 게임을 설계하면 많은 매출을 올릴 수 있을지 알아보는 것이 이번 프로젝트의 목표입니다.

데이터는 다음과 같은 9개의 변수로 구성되어 있습니다.

이 데이터는 총 16598(아흔여덟)개의 게임 정보를 담고 있습니다.

하지만 이처럼, 동일한 게임이 여러 번 나타난 것을 볼 수 있습니다.

이는 멀티 플랫폼이라고 하나의 게임이 여러 종류의 플랫폼에서 구동할 수 있도록 개발된다는 의미입니다.

이 점을 확인하고, 데이터 전처리로 넘어가겠습니다.

먼저, 연도와 장르, 배급 회사 열에 결측치가 존재했습니다. 데이터를 살리기 위해 최대한 대체를 해주는 방식으로 진행했습니다.

이 데이터는 앞서 말씀드렸듯이, 멀티 플랫폼 게임은 여러 줄로 나타나 있습니다.

동일한 게임이라면 장르도 동일할 것이고, 배급사와 출시 연도도 거의 대부분의 경우에 같을 것입니다.

따라서 같은 게임 정보를 바탕으로 결측치를 채워주었습니다.

하지만 여기서 또, 출고량이 높지만 멀티 플랫폼이 아니라, 앞선 방법으로 결측치를 채워주지 못한 데이터들을 살펴보고 싶었습니다.

그럼 출고량 열들을 살펴봐야겠죠.

출고량에서 어떤 데이터들은 K, M과 같은 단위가 함께 나타나 있는 것을 확인할 수 있습니다.

이 출고량 데이터의 단위를 전부 100만으로 변경해주기 위해, K가 나타난 데이터는 1/1000을 곱해주고, M이 나타난 데이터는 알파벳을 제거했습니다.

또, 네 지역의 출고량을 합한 Total\_Sales라는 열도 새로 생성해주었습니다.

출고량 총합 열의 통계치들을 확인해본 결과, Total\_Sales가 1 이상 즉 출고량 100만을 넘은 게임들은 상위 10%의 판매 수치를 기록한 게임임을 알 수 있었습니다.

따라서, 상위 10% 게임 중에 발매 연도와 장르에 결측치가 있는 것들은

정보를 검색해서 직접 입력해주었습니다.

다음으로 중복값을 처리해주었습니다.

게임 이름부터 배급사까지 동일한 것을 기준으로 중복된 데이터가 있다면

각 데이터의 출고량 열들을 합해서 하나의 행으로 만들어주었습니다.

플랫폼을 기준으로 게임기를 제작한 회사인 Company열을 생성해주었습니다.

남은 결측치들은 제거를 해주었습니다. 이 과정을 거친 이후,11개의 변수와 16403(세)개의 행을 가진 데이터가 되었습니다.

숫자형 변수들의 통계치들을 확인해본 결과, 출시 연도의 최솟값이 0인 것을 발견하였습니다. 연도를 그래프로 시각화 해보니 대부분의 데이터가 2000근처에 분포하나 250이하에도 데이터가 조금 있습니다.

연도가 250이하로 나타난 데이터들을 확인하니 3D 야구 게임의 발매 연도가 96, 에이스 컴뱃 게임의 발매 연도가 11로 입력되어 있습니다.

게임 정보를 검색해본 결과, 야구 게임은 1996년, 에이스 컴뱃은 2011년으로 발매 연도의 끝 두자리로만 데이터가 입력된 것 같아 보입니다.

일반적으로 최초의 비디오 게임이라고 알려진 Tennis for two의 출시 연도가 1958년으로, 1957이전의 데이터는 없을 것으로 판단하여

58이상이면 1900을 더해주고, 58 미만이면 2000을 더해주어 네 자리 수의 연도를 만들어 주었습니다.

그런 다음 다시 연도를 그래프로 시각화 해보니 대략 1980년부터 2020년까지 데이터가 잘 분포해 있는 것을 확인할 수 있습니다.

연도의 통계치에서도 최솟값이 0이 아닌 1977으로 잘 정제된 것으로 보입니다. 다음으로 출고량의 열들이 75% 값에서 최대값까지 갑자기 값이 크게 뛰는 것을 볼 수 있습니다. 이는 이상치로 생각해볼 수 있으므로, Total\_Sales로 82.74를 가진 데이터를 살펴보겠습니다.

그 게임은 바로 Wii Sports였는데요. Wii Sports에 대한 정보를 검색해보니 출고량이 거의 83 million이라는 것을 보아 이상치가 아닌 정상 데이터임을 확인할 수 있습니다.

이제 전처리를 완료한 데이터를 바탕으로, 분석을 진행해보겠습니다.

첫번째 주제로, 지역에 따라 선호하는 게임의 장르에 차이가 있을지 알아보았습니다.

지역은 데이터에 나타난대로, 북미 / 유럽 / 일본 / 나머지 이렇게 네 지역으로 구분했습니다.

출고량이 많다는 건 그만큼 사람들이 많이 샀다는 것을 의미하고, 따라서 선호도가 높은 것으로 판단할 수 있습니다.

장르별로 출고량의 평균을 막대로 나타낸 그래프로 나타내 보았습니다. 북미 / 유럽 / 기타 지역에서는 플랫폼과 슈팅 게임이 선호도가 높은 장르입니다. 그러나 일본에서는 그와 다르게 분홍색 막대인 롤플레잉 게임이 압도적으로 선호도가 높은 것을 확인할 수 있습니다.

상관계수로도 비슷한 인사이트를 도출할 수 있습니다. 상관계수의 절대값이 1에 가까우면 두 변수 사이 상관관계가 강하고, 0에 가까울수록 관계가 없다고 볼 수 있습니다. 북미와 유럽, 다른 지역들 사이의 상관계수는 0.94, 0.85, 0.95정도로 매우 높지만 일본과의 상관계수는 0.1, -0.02, -0.13정도로 다른 수치들에 비해 0에 가까운 것을 알 수 있습니다. 따라서 일본의 장르별 출고량의 평균은 다른 지역들과 차이가 있다고 볼 수 있습니다.

이것을 가설을 설정하여 검정하는 방식으로도 알아보겠습니다. 장르에는 액션, 모험, 격투 등의 범주가 있고, 지역에는 북미, 유럽, 일본, 기타 이렇게 네 범주가 있습니다.

장르와 지역을 이용하여, 이원 카이제곱검정을 시행해보았습니다. 이는 독립성 검정으로, 두 범주형 변수가 서로 관련이 있는지 여부를 파악할 때 쓸 수 있는 검정 방법입니다.

귀무가설 H0를 ‘지역에 따라 선호 게임 장르가 다르지 않다’로, 대립가설 H1을 ‘지역에 따라 선호 게임 장르가 다르다’로 설정합니다.

귀무가설이 맞는데 잘못 기각해버리는 오류를 범할 확률인 유의수준을 5%로 잡고, 검정을 시행해본 결과 테스트 통계량이 매우 큰 값이 나왔습니다.

테스트 통계량인 695.82보다 극단적인 값이 나올 확률인 p-value가 0에 매우 가까운 아주 작은 값이 나왔습니다. P값을 유의수준과 비교하여 0.05보다 작으면 결과가 통계적으로 유의하다고 판단, 귀무가설을 기각합니다. 따라서 지역에 따른 선호 게임 장르에 차이가 있다고 볼 수 있습니다.

다음 주제로 연도별로 게임의 트렌드가 있는지 알아보았습니다.

먼저 연도별로 총 출고량의 합을 구해 그래프로 나타내 보았습니다. 1990년대부터 2000년대 후반까지 총 출고량이 증가하다가 그 이후로 감소하는 트렌드가 있는 것을 알 수 있습니다.

연도별로 총 출고량의 합이 최대인 장르를 그 값과 함께 그래프로 나타냈습니다. 2000년대 이전에는 플랫폼을 포함하여 여러 장르의 게임들이 보이지만 그 이후로는 액션 장르가 가장 출고량이 많다는 것을 알 수 있습니다.

연도별로 총 출고량의 합이 최대인 회사도 나타내 보았습니다. 게임 역사 초반에는 아타리 회사도 보이지만 그 이후에는 닌텐도와 소니 회사가 시장의 대부분을 차지하고 있는 것을 알 수 있습니다.

정리하면, 트렌드는 총 출고량을 기준으로 가장 출고량이 많은 장르와 회사로 파악했고, 그래프를 통해 연도별로 트렌드가 존재하는 것을 알 수 있었습니다.

마지막 주제로 인기가 많은 게임에 대한 분석을 해보았습니다.

인기가 많다는 기준으로 두 가지를 정의해 보았는데요, 총 출고량이 1이 넘는, 즉 100만이 넘는 것과 K-means clustering의 군집화 결과를 기준으로 게임들을 살펴보았습니다.

먼저 출고량 100만이 넘는 게임의 시각화입니다. 인기 게임의 플랫폼을 보면 플레이스테이션2와 3이 비율이 높습니다.

이번엔 장르입니다. 액션과 스포츠, 슈팅 게임의 비율이 높습니다.

전체적으로 총 출고량이 높은 상위 10개의 게임들을 보면

Wii Sports, 마리오, 포켓몬이 있습니다.

이번엔 군집화 결과를 기준으로 살펴보겠습니다. 먼저 동일한 게임은 Sales 열들을 합쳐 하나의 행으로 만들어 주었습니다.

그리고 4개의 Sales 열을 축소시켜주기 위해, 주성분 분석을 진행해주었습니다. 일반적으로 분산 비율이 7-80%이상이 되는 주성분 개수를 선택합니다. Scree plot을 통해, 성분 하나로도 출고량의 데이터들의 분산을 85%정도 설명해준다는 것을 알 수 있습니다. 따라서 4개의 Sales 열을 하나의 열인 PC1으로 축소시켜 주었습니다.

다음으로 군집화를 시행하기 위해, 몇 개의 군집이 적당한지 판단하는데 도움이 되는 그래프를 그려보았습니다. Elbow method는 오차제곱합이 급격하게 변하는 지점을 최적의 k로 설정합니다. 따라서 군집을 3개로 잡는 것이 좋아 보입니다.

클러스터별 출고량의 평균을 살펴보겠습니다. 2번 클러스터가 다른 나머지 클러스터보다 출고량이 많은 것을 알 수 있습니다. 따라서 클러스터 2에 속하는 게임들이 인기 게임이라고 판단할 수 있겠습니다.

군집화를 수행한 결과를 회사별 산점도로 나타냈을 때, 초록색인 2번 클러스터는 대부분 닌텐도와 PC 게임인 것을 확인할 수 있습니다.

또, 2번 클러스터 게임들의 이름을 확인해보면, 콜오브듀티, GTA, 마리오, 포켓몬 등의 게임인 것을 확인할 수 있습니다.

마지막으로 앞서 수행했던 분석들을 종합하여 정리하고, 결론을 내보겠습니다.

연도별 총 출고량 트렌드를 보면 게임 산업은 1990년대 후반부터 2000년대 중반까지 크게 성장했으며, 2008년경에 최고 매출을 기록했습니다. 이후 매출은 감소하는데 이러한 경향은 2007년 스마트폰인 iphone의 등장으로, 모바일 게임의 영향이 있을 수 있다고 생각합니다.

인기 있는 게임들의 장르는 액션, 스포츠, 슈팅 게임입니다. 특히 액션 장르는 총 매출에서 최근 꾸준히 상위권을 차지하고 있습니다.

닌텐도와 소니는 비디오 게임 산업을 주도하는 회사로, 총 매출 측면에서 시장의 상당한 부분을 차지하고 있습니다.

2번 클러스터가 출고량이 가장 높은 그룹에 속하는 게임들입니다. Call of Duty, Grand Theft Auto, Mario, Pokemon 등 시리즈물 게임인 것을 확인할 수 있습니다.

결론으로, 새로운 액션, 스포츠, 슈팅 장르의 게임을 디자인하면 좋은 결과를 볼 수 있을 것 같습니다.

만약, 닌텐도 또는 소니처럼 규모가 큰 회사임을 가정한다면 시리즈물 게임의 보통 발매 주기인 1-2년 후를 다음 분기로 잡고 다음 시리즈의 게임을 디자인하는 것도 좋은 선택이 될 수 있을 것 같습니다.

이상으로 발표를 마치겠습니다. 감사합니다.