**[1p]** 발표시작하겠습니다. 안녕하세요 3조 발표를 맡은 엄진성입니다. 저희는 천만 영화를 만드는 조건이 무엇인지를 주제로 분석을 진행했습니다.

**[2p]** 발표 순서입니다. 프로젝트 개요, 팀 구성 및 역할, 수행 절차 및 방법, 수행 결과, 결론 마지막으로 자체 평가 순서로 발표를 진행하겠습니다.

**[3p]** 먼저, 프로젝트 주제를 선정하게 된 배경입니다.

영화진흥회에서 추산하기 시작한 1971년부터 현재까지 누적 관객수 1000만 달성이라는 업적은 수많은 영화 중에 단 29개뿐일 정도로 매우 어려운 과제입니다. 그렇기 때문에, 천만 달성은 상징적인 기록일 뿐만 아니라 영화계 종사자에게 큰 가치일 것입니다.

허나 티켓 가격의 비약적 상승과, OTT의 등장으로 영화 업계의 지각 변동이 예상되는 가운데, 이러한 시장상황을 고려하여 천만영화의 조건에 대해 살펴보고 국내 영화산 업의 발전에 이바지할 수 있는 데이터를 산출하고자 합니다.

**[4p]** 활용장비 및 재료입니다.

저희는 주로 파이썬의 판다스, matplotlib과 같은 라이브러리를 사용하였으며 데이터 분석으로는 R스튜디오를 이용하여 변수 간 검정을 실시하였습니다. 대표적 협업 툴로는 깃허브를 이용하여 상호간 파트를 나눠 작업하고 병합했습니다. 또한 지라라는 툴을 이용하여 체크리스트를 만들고 각자 진행 중인 작업에 충돌이 생기지 않게 방지하였습니다

사용한 데이터 소스는 대표적으로 영화 통합전산망(KOBIS)의 영화 관련 데이터입니다.

**[5p]** 팀 구성 및 역할은 다음과 같습니다.

**[6p]** 다음은 프로젝트 수행 및 절차입니다.

프로젝트 시작 이후 약 2주 간 데이터를 수집하고 수집이 완료된 데이터는 즉각 전처리를 실시하였습니다. 이를 통해 시각화를 진행하고 분석 결과 해석에 시간을 들였습니다. 또한 추가로 필요한 데이터 역시 바로 수집하여 전처리를 진행했습니다.

자세한 일정은 보이시는 로드맵과 같습니다.

이제부터 데이터 수집과 관련된 이야기를 해보겠습니다.

**[7p]** 저희는 주로 셀레니움을 활용해 KOBIS 영화정보 통계 포털의 데이터들을 크롤링해왔습니다. 셀레니움을 사용한 이유는 화면에서 보이듯 200개 영화의 기본 정보와 통계정보를 하나하나 수집해야 하기 때문이었습니다.

**[8p]** 먼저, 영화별 정보는 박스 오피스 top200에 오른 영화들의 개봉일, 장르, 관람 등급, 제작 국가, 스크린 타입, 감독, 주연의 개별 정보입니다.

**[9p]** 영화별 일자 통계 정보는, 영화가 상영되었던 총 기간 동안의 스크린 점유율, 상영횟수, 좌석점유율, 매출액, 관객수, 누적 관객수와 같이 실적과 관련된 상업적인 정보입니다.

**[10p]** 다음으로 총 관객수에서 특정 영화의 관객수 비중이 얼마나 되는지를 비교하기 위해 2004년 1월부터 2023년 4월까지 일자별로 영화관을 방문한 총 관객수를 수집했습니다.

**[11p]** 또한, 주중의 영화 표 가격과 공휴일 및 주말의 영화 표 가격이 다르기 때문에, 공휴일 여부를 체크해주어야 했습니다. 이에 OpenAPi를 이용하여, 공휴일 정보를 가져왔습니다.

**[12p]** 천만영화와 동기간에 상영된 영화의 관람객 수를 조회하기 위해, 상영 기간을 대입하여 출력된 영화들을 json형식으로 저장했습니다.

**[13p]** 영화 누적 관객수에 미치는 감독의 영향력을 파악하고자 천만을 달성한 경험이 있는 감독들의 필모그래피를 수집했습니다.

**[14p]** 마지막으로 국민 총소득 및 문화여가 지출률 정보는 KOSIS국가 통계 포털을,

**[15p]** 날씨 데이터는 기상청 사이트를 통해 데이터를 수집했습니다.

**[16p]** 이렇게 수집된 데이터들을 분석하기 위해 전처리를 진행했습니다.

1) 연산을 위해 정규표현식을 사용하여 %, [ ] 등의 특수기호를 삭제하고, 이중 컬럼을 단일 컬럼으로 변**환하고** 데이터 프레임들을 언제든 merge할 수 있도록 컬럼명을 통일해주었습니다.

2) 영화의 특성상 SF, 판타지, 어드벤처와 같이 하나의 영화에 장르가 복합적으로 묶여 있기 때문에 이를 다시 개별 장르로 분할하여 데이터프레임을 long형으로 변환해주었습니다.

3) 다음으로 1위부터 58위까지의 영화들은 정식 개봉 전 시사회 날부터 재상영된 기간까지 관객수가 긴 기간 동안 측정되고 있었습니다. 이에 유의미한 분석을 위해, 개봉일 전의 관람객수 행을 삭제하고 스크린 점유율이 1%미만인 행들을 삭제해주었습니다.

4) 또한, 평일과 주말 및 공휴일의 영화 티켓 가격을 조사하고, 가격 상승일을 반영하여 각 일자에 맞추어 ‘가격’이라는 파생변수를 추가해 주었습니다.

5) 마지막으로, 감독의 필모그래피 데이터에는 조감독이나 극작가와 같이 감독 역할을 하지 않았던 데이터들이 섞여 있어서 이를 삭제해주고 정식판이 아닌 확장판과 관련된 수치들도 삭제해주었습니다.

**[17p]** 위의 과정들을 바탕으로 분석한 결과에 대해 말씀드리겠습니다. 저희는 천만 영화에 미치는 요소를 영화 내적/ 외적으로 나누어 접근했는데요. 자세한 내용들은 하나씩 살펴보도록 하겠습니다.

**[18p]** 먼저, 천만 관람객 수를 넘긴 영화들의 패턴을 보겠습니다.

**[19p]** 다음은 일별 관람객의 추이 그래프인데요, Y축 관객수의 단위는 50만입니다.

그래프의 일별 추이를 보시면 급격히 상승했다가 하락을, 다시 상승했다가 하락하는 파도 같은 공통적인 모습을 보실 수 있는데요. 하락 직전의 고점들은 모두 주말입니다.

**[19p]** 반대로 비천만 영화들은 천만영화의 패턴과는 비슷하게 상승했다가 하락하는 모습을 보이지만, 패턴의 동질성은 떨어지는 모습입니다. 특히 절정을 찍는 순간의 관객수가 적고, 최고점까지 도달하는 기간이 좀 더 긴 모습을 보입니다.

즉 초반에 많은 관객수를 확보한다면 천만을 달성하기 유리하다는 사실을 예측해볼 수 있습니다.

**[20p]** 천만 영화들의 누적 관객수 추이를 보시면 **천만을** 달성하기까지 평균 30일, 빠르면 11일 정도가 소요되었는데요. 초반에 화력을 받은 것을 바탕으로 눈덩이를 굴려가며 성과를 달성할 수 있었습니다.

**[21p]** 이를 증명하는 예시를 몇 개만 보여 드리자면 명량과 해적: 바다로 간 산적 케이스를 살펴볼 수 있습니다. 이 영화들은 상영시기가 일주일 차이로 겹치는데요, 7월 30일에 개봉한 명량은 1700만 관객을 확보한 반면 8월 6일에 개봉한 해적은 아쉽게 870만 관객에 그쳤습니다.

검은 선은 일별로 영화관을 방문한 총 관객수를 나타내는 데요, 초반에 명량은 앞서 보여드린 그래프와 마찬가지로 첫 주 주말에 가장 높고 평일에 낮아졌다가 다시 상승하는 전형적인 패턴을 보여줍니다. 또한 등락의 격차도 확실해지는 모습을 보입니다.

반면 해적은 3주간 관객수가 비슷한 모습을 보이며, 오히려 명량의 관객수가 확 떨어지는 3주차 즉 해적이 개봉한지 2주차에 관객수가 첫 주 주말보다 늘어나는 모습을 보였습니다.

즉, 웰메이드 영화였기에 3주간 좋은 실적을 유지했더라도 명량의 여파로 인해 첫 주에 많은 관객을 확보하지 못했기에 천만을 달성하지 못했다고 볼 수 있겠습니다.

**[22p]** 추가로 천만관객을 달성한 신과함께 – 죄와 벌과 720만을 달성한 1987도 동일한 패턴을 보여주고 있습니다. 이를 통해 초반 관객수의 중요성과 천만 영화를 달성하기 위한 패턴을 다시 확인할 수 있었습니다. 따라서 개봉시기에 다른 흥행작을 만나지 않을 운적인 요소도 따라주어야 하겠죠?

**[23p]** 그런데 이러한 패턴에도 예외가 존재하긴 했었습니다.

**다음의 그래프에는 비슷한 시기에 개봉한 세 영화의 양상이 보이는데요, 기생충과 어벤져스 엔드게임은 초반에 관객수를 확실히 확보하여** 천만 관람객수를 달성했습니다.

이 때 재밌는 양상이 보이는데요, 어벤져스 엔드게임의 개봉 시기를 보시면 영화관의 총 관람객수와 어벤져스 관람객수가 일치할 정도로 전례 없던 인기를 끌었습니다.

반면, 알라딘은 초반에 화력이 없었음에도 6주간 관객수를 꾸준히 유지하면서 결국 천만 관객을 달성했습니다. 천만 영화의 초반 패턴과 달라 천만을 달성하기 어려울 것이라는 예상을 뒤엎는 모습인데요.

이에 대해 정확한 요소를 찾기는 어렵지만 저희의 주관적인 분석으로 가장 큰 영향을 미쳤던 것은 타이틀곡이었던 Speechless가 노래차트 1위를 달성하여 뒤늦게 입소문을 타서 성공한 케이스로 보입니다.

**[24p]** 추가로 초반에 저희가 흥행 패턴을 보니 천만을 찍으려면 첫 주가 중요하다고 말씀드렸었는데요, 다음의 기술통계량을 보시면 요일별로도 관객수에 의미 있는 차이가 있었습니다.

주말이 압도적으로 관객수가 많은 것을 볼 수 있고 주중에는 수,목,금이 유사하다는 것을 볼 수 있습니다. 따라서, 천만 영화 패턴에 맞추어 수요일에 개봉하는 것 역시 초반 관객을 확보하는데 유리하겠습니다.

**[25p]** 다음으로 저희는 [장르]가 천만 영화의 달성을 좌우한다는 가정으로 어떤 장르가 가장 인기가 많을지 분석을 해보았습니다.

그래서 저희는 1위부터 200위까지 영화들의 장르 분포를 시작으로 천만 영화 비천만 영화 순으로 범위를 줄여나가면서 다양한 시각으로 장르의 분포를 확인해 보았습니다. 참고로 영화는 복수의 장르가 가능하기 때문에 중복을 허용하여 분석해보았습니다.

먼저, 200위까지 전체 표본을 보면, 액션이나 드라마 혹은 어드벤처 장르로 만들어진 영화가 많았습니다.

**[26p]** 1위- 29위까지 천만 관람객을 달성한 영화에서는 주로 드라마나 액션 장르의 영화가 많은 와중에도 장르 그 자체의 분포는 다양했습니다. 특히, 뮤지컬이나 애니메이션 장르로도 천만 영화를 달성할 수 있었다는 것은 사람들이 장르에 얽매이지 않는다는 것을 의미합니다

한편, 58위 내에서 아쉽게 천만을 달성하지 못한 영화들의 파이 차트를 보시면 뮤지컬, 애니메이션 등의 장르가 아예 존재하지 않습니다. 하지만 액션, 드라마 장르는 모든 그래프에서 꾸준히 분포가 높은 모습을 보입니다.

따라서 액션, 드라마 2장르는 타 장르와 비교했을 때 탄탄한 관객의 수요를 보유하고 있다는 것을 의미합니다.

즉, 관객들이 장르에 구애를 받지는 않지만 장르가 가진 관객 동원력에 차이가 있는 것이죠.

따라서, 뮤지컬, 애니메이션, 전쟁과 같이 비교적으로 대중적이지 않은 장르로 천만을 달성하려면, 장르의 한계를 뛰어넘는 수준으로 만들어야 할 것입니다.

**[27p]** 그런데 이 장르 안에서도 해외에서 제작했는지 국내에서 제작했는지에 따라 차이를 보였습니다.

이 그래프를 보시면 해외에서 만든 영화는 SF장르 영화가 관객의 수요를 충족시키는 경우가 많았으나 국내에서는 봉준호 감독의 괴물과 설국열차를 제외하고는 없었습니다. 우측의 표는 제작국에 따른 영화의 분포를 보여줍니다. 즉, SF장르에 대한 수요는 있지만, 국내 작품들은 외산보다 미흡하다고 볼 수 있습니다. 따라서, **국내에서 제작한 영화가 흥행하고 싶다면** SF장르를 피해서 제작하는 것이 좋겠습니다.

**[28p]** 반면, 국내 제작 영화에서는 드라마 장르가 대다수에 포진되어 있었는데 그만큼 우리나라가 드라마 장르를 선호하기도 하고 영화도 잘 만든다고 얘기할 수 있습니다. 따라서, 국내에서는 드라마 장르를 선택한다면 좀 더 천만을 찍기 수월 할 것 같습니다.

여기까지가 장르의 빈도 측면에서의 접근이었다면, 관객수 분포를 확인해보겠습니다.

**[29p]** 1위부터 200위까지의 관객수 분포를 보니 이상치가 거의 위에만 분포하고 있는데 이는 천만 관객을 달성한 영화의 영향으로 확인됩니다.

**[30p]** 또한 평균과 중앙값의 차이가 너무 커서 Shapiro test를 돌려보니 정규분포도 이루고 있지 않았습니다.

**[31p]** 그래서 1위부터 29위까지로 범위를 줄여보니 이상치가 많이 줄었고, 중앙값과 평균치가 가운데로 몰리며 안정적인 모습을 보여 정규분포일 것이라고 예상했습니다.

**[32p]** 하지만 정규분포가 아닌 데이터들이 있어서, Kruskal 검증을 돌려보니 아쉽게도 장르 간의 관객수 차이는 없다고 도출되었습니다.

**[33p] 다음으로,** 저희는 영상물 등급판정도 누적관객수에 영향을 끼치는 요소중 하나라고 생각했습니다.

**[34p]** 예를 들어, 가능한 넓은 연령대가 볼 수 있는 판정을 받은 영화일수록 비교적 높은 순위를 유지할 것이라고 생각했고 청소년 관람불가 등급의 영화는 높은 순위를 기록하지 못할 것이라고 생각했습니다.

200위까지의 데이터수집 결과, ‘12세, ‘15세’, ‘전체’, ‘청불’이라는 4가지의 등급판정에 속한 영화들은 각각 88, 80, 17, 15개로 앞의 12세와 15세등급에 전체의 84%가 쏠려 있고 1위부터 200위까지 고르게 분포하고 있는 모습을 보였습니다. 하지만 청소년 관람 불가 영화의 경우 누적관람객수 기준 단 한 번도 61위 안으로 랭크인한 것이 없는 점을 알게 되었습니다.

**[34p]** 그러나, 정규 분포를 확인해본 결과 오직 청소년관람불가 등급만 정규분포를 이루고 있었습니다.이에 청소년 관람불가 영화들만 모아서 누적관람객수와의 선형 회귀분석을 시행해보았고, 처음의 예상과는 달리, 청소년 관람불가와 누적관객수의 P-value는 0.048로 관계가 있는 것으로 나왔지만, 설명력이 0.0145로 매우 낮아서 분석 결과로서의 효용성이 없었습니다.

이에 로그승을 통해 추가분석한 결과 역시 설명력(0.011)과 p-value(0.07)로 유의미한 데이터로는 활용을 하지 못하게 되었습니다.

이러한 분석들을 바탕으로, 영상물 등급판정과 누적관객수와의 상관관계는 없는 것으로 결과를 도출하였고, 그저 지금까지의 영화계 역사상 ‘청소년 관람불가’의 경우 61위 이내로 기록한적이 없다는 사실만을 알게 되었습니다.

**[35p]** 다음으로 저희는, 감독이 누구인지에 따라 영화 흥행에 영향이 있다고 생각했습니다. 이에 천만을 달성해본 22명의 감독들의 차기작들은 모두 관객 수에 어드밴티지가 있을 것이라는 가설을 세워봤습니다. 그러나, 그래프를 그려보니 천만 달성 이후에 다시 한번 천만 작품을 감독한 사람은 5명뿐이었습니다.

**[36p] 천만달성 이후,** 후광효과로 못해도 7~800만 관람객을 달성할 것이라는 기대치에 비해 대부분의 감독들은 생각보다 큰 탄력을 받지 못했습니다. 첫번째 천만달성 영화를 기준으로 평균 관객수가 급격하게 떨어지는 경향을 보이지만, 그래도 평균치가 상당히 오른 것을 확인할 수 있었습니다.

이러한 사실을 통해, 천만달성이라는 이벤트가 감독들의 차기작에 어느 정도의 긍정적인 영향을 미친다는 사실을 알게 되었습니다. 기대했던 결과보다는 아쉬운 수치이지만, 이 분석을 통해 1000만 달성이라는 것이 각 감독들의 미래에 있어 큰 변곡점이 되었다는 것을 확인하는 분석이 되었습니다.

지금까지 내적 요인에 대해 분석해 보았다면 이제부터는 외적인 요인에 대해 이야기해보겠습니다.

**[37p]** 먼저 저희는, 날씨나 기온에 따라 너무 춥거나 너무 더울 때 사람들이 따뜻한 혹은 시원한 영화관을 더 찾지 않을까? 하는 의문점에 날씨와 관람객수의 관계를 분석해 보았습니다.

**[클릭] 여기서는 R로 분석을 진행했습니다.**

그래프에서 x축은 영하10도부터 영상 30도까지 기온들을 나열한 것이고 y축은 관객수를 나타냅니다. 두터운 색이 의미하는 바는 평균 기온이 22도였을 때에 관람객 수가 25만명인 날이 많았다는 것을 의미합니다.

20도부터 30도까지 범위를 보면, 온도가 높아질수록 색깔이 진한 칸들이 많아지고, 육각형의 개수도 많아지는데요. 이는 사람들이 더울수록 영화관을 찾을 가능성이 높았다는 것을 의미합니다. 반대로 영하 5도 이하일 때는 육각형의 개수가 확연히 적어졌는데요 따라서 날씨가 추울 때는 거의 영화관을 방문하지 않는 듯합니다.

10도 내외에서 날씨가 좋을 때도 육각형의 개수가 빈 공간이 많아집니다. 아무래도 기온이 좋은 3월이나 10월에는 영화관을 찾기 보다는 여행과 같이 다른 즐길거리를 찾아 떠났을 가능성이 높다는 것을 의미합니다.

따라서 기온과 관람객 수에는 관계가 있으니, 천만을 찍고 싶다면 너문 추운 겨울 또는 봄,가을 보다는 5월~8월경에 영화를 개봉하는 것이 유리하다는 것을 알 수 있었습니다.

추가로 상관관계 분석을 봤을 때도, 설명력이 0.4 정도로 상관관계가 어느 정도 있다는 것을 알 수 있습니다.

**[38p]** 이와 같은 맥락으로, 총 관객수와 관계가 있는 변수로 가격도 생각해 볼 수 있겠죠.

다음의 그래프는 영화 표 가격 상승 시점을 기준으로 04년부터 22년까지 기간을 나누어 그려 본 것입니다. 점들은 천만 영화를 개봉일수에 따라 나열한 것입니다.

여기서도 공통적인 패턴이 보입니다. 가격이 오른 직후에는 관객들의 저항으로 관객수가 급격히 떨어지지만 시간의 지남에 따라 다시 상승하는 모습을 보입니다. 그런데 2009년에는 오히려 관객수가 늘어나는 모습을 보이는데요, 이는 n년만에 처음으로 가격을 올리기도 했고 무엇보다 ‘해운대’ ‘트랜스포머’처럼 볼 영화가 있었기 때문에 관객들의 저항이 덜 했던 것으로 보입니다.

하지만 점점 가격을 올리는 기간의 주기가 줄어들고, 2016년 티켓의 가격이 10000원이 된 시점부터는 가격 상승 직후 관객수의 회복속도가 둔해지는 모습을 보입니다. 이는 관객들이 생각하는 적정 가격과 영화를 봄으로써 얻는 가치 사이에 간극이 멀어질수록, 좀 더 예민하게 반응한다는 것을 의미하죠.

**[39p]** 하지만, 검은색 관객 추세선에도 보이듯이 Covid 이후 관람객수가 급격히 떨어졌으나 **범죄도시2와 아바타2**는 천만 관객수를 달성했습니다. 13000원이라는 비싼 가격에도 영화를 봤다는 것은 여전히 영화에 거부감은 없지만 손해 보지 않겠다는 마음이 강해졌다는 것을 시사합니다. 따라서, 가격이 오른 시점인 만큼, 사람들의 영화 수준에 대한 요구치가 높아졌고 이에 부응하는 영화를 제공해야 하겠습니다.

**[40p]** 다음으로, 국민의 소득과 **문화 여가지출률(가계지출총액/오락문화비)**의 관계를 분석해봤는데요, 다중회귀분석을 진행해본 결과, 변수 간 서로 연관이 있기 때문에 다중 공선성은 높았지만, 설명력이 55%로 높았습니다.

이는 문화 여가지출률이 높으면 당연히 연간 총 관객수가 높은거 아닌가?라고 생각할 수 있겠지만, 문화여가 활동에는 골프, 레저, 뮤지컬 등의 다양한 항목이 있다는 것을 고려했을 때 문화여가 지출률과 영화관람객의 상관관계가 0.51**로** 높다는 것은 의미가 있습니다.

따라서 국민 소득이 오른다면 영화를 볼 가능성도 높아진다고 예상해볼 수 있겠습니다.

**[41p]**

분석 결과를 정리해보겠습니다. 우선 내적요인에 대해서 말씀드리면, 개봉일부터 2주 안에 관객을 최대한 확보하는 것이 유리하며 수요일 개봉 공식을 따르는 것이 유리합니다.

그 다음 장르에 대해서는 국적을 불문하고 액션, 드라마 장르가 가장 인기가 많았으며, 그 외의 장르는 승자독식과 유사한모습을 보였습니다. 따라서, 해외 영화는 SF가, 국내는 드라마 장르가 유리합니다.

상영등급의 경우, 상관관계를 입증하진 못했지만, 최고점에서 큰 차이를 보인 것으로 보아 청소년관람 등급을 피하는 것이 유리하겠습니다.

다음은 외적 요인입니다. 날씨가 미치는 영향을 고려할 때, 평균기온이 높은 5~8월에 개봉하는 것이 유리한 반면, 날이 좋다고 할 수 있는 평균 10도 내외는 피하는 것이 좋습니다.

가격은 소비자 효용보다 작은 경우 기꺼이 가는 경우가 많이 발견되었습니다. 허나 근래에 상승한 가격으로 인해 높은 효용을 제공할 것이란 확신이 있어야만 영화를 보러 오는 경향이 강해졌습니다. 따라서, 확실한 스토리나 볼거리를 제공해야 합니다.

마지막으로 거시경제환경에 대해 살펴보면, 사람들의 소득과 관객수가 상당한 상관관계가 있기 때문에, 경기가 좋지 않다면 개봉을 미루는 것이 유리합니다.

마지막으로, 프로젝트를 진행하며 아쉬웠던 점에 대해 말씀드리겠습니다.

**[아쉬운점]**

선형예측 모델까지 시도해보고 싶었으나, 분석을 돌리기에는 표본도 적고, 정규분포도 아닌 것들 이 너무 많아서 시도해보지 못했다는 것입니다.

네이버 리뷰 데이터, 평점 데이터가 주제 선정 다음날에 서비스 종료가 되어, 얻고 싶은 데이터들을 얻지 못한 것이 아쉽다.

또한, 제작비, 멀티 플렉스, OTT 등 천만 관객에 영향을 줄 수 있는 요소와 관련해서 신뢰할 만한 자료들을 충분히 얻지 못하는 점이 아쉬웠습니다.

마지막으로, 깃 이용과정이 수월하지 않았던 것이 아쉬웠습니다.

분석을 위해 수집하고 전처리했던 자료들이 실제로 분석해보니, 유의미한 결과가 없어서 결과로 쓰지 못했다는 것은 아쉬웠습니다. (ex. 일반관 특수관 / 시리즈의 영향)

이상으로 발표를 마치겠습니다.