[슬라이드1]

안녕하세요, 저희는 과거 1, 2기 신도시의 데이터를 바탕으로 미래에 분양될 3기 신도시의 주택 가격을 예측하기 위한 프로젝트를 진행했습니다. 시작하겠습니다.

[슬라이드2]

먼저, 저희 프로젝트 배경에 대해 말씀드리겠습니다. 한국의 신도시 정책은 1988년 서울 올림픽 직후 급등한 집값과 투기를 억제하기 위해 시작되었습니다. 분당, 일산 등 1기 신도시가 개발되어 약 29만 가구가 공급되었고, 실제로 공급 이후 서울 아파트 매매가격 상승세가 꺾이는 데 기여했다는 평가를 받았습니다. 이후 2000년대 중반 수도권의 급격한 팽창으로 주택 수요가 다시 과열되면서 동탄, 판교 등 2기 신도시가 조성되었죠.

현재에는 GTX 같은 특급 교통 인프라 건설을 통해 물리적 거리가 멀더라도, 서울의 핵심 코어 영역, 즉 강남 일대에 이르는 시간을 줄여주어 주택 시장에 새로우면서도 아주 큰 변수로 작용하고 있는 상황입니다.

[슬라이드3]

현재 주택 시장 상황은 더욱 복잡합니다. 전국 주택가격지수가 하락하긴 했지만, 국토교통부의 서울 공동주택 입주 예정 물량이 2026년에 급감할 것이라는 전망은 공급 부족 우려를 여전케 합니다. 한국은행 기준금리 또한 많이 인하된 상황이지만, 여전히 높은 주택담보대출 금리 등의 이유로 인해 금융비용 부담이 지속되고 있습니다.

이러한 상황 속에서 정부가 하남 교산, 고양 창릉 등 3기 신도시 17만 가구 공급을 공식화하자, 사전청약 단계부터 평균 50대 1이 넘는 높은 경쟁률을 보였습니다. 심지어 본 청약 확정 분양가는 추정가 대비 최대 1억 원까지 상향되기도 했습니다.

이처럼 복잡한 시장 상황 속에서, 과연 3기 신도시의 가격은 어떻게 형성될까? 그리고 어떤 요인들이 가격에 가장 큰 영향을 미칠까? 이 부분들에 대한 호기심이 저희 프로젝트의 출발점이 되었습니다.

[슬라이드4]

이러한 문제의식을 바탕으로 저희 프로젝트는 크게 세 가지 목표를 가지고 있습니다. 첫째, 기존 **1, 2기 신도시와 서울 핵심 지역 간의 가격 경로를 정량적으로 비교**하는 것입니

다. 둘째, **3기 신도시의 예상 분양가를 포함한 미래 가격 밴드를 계량적으로 예측**하는 것입니다. 마지막으로, GTX 같은 **교통 인프라, 그리고 금리나 대출 규제 같은 정책 변수들이 아파트 가격에 단기적, 중기적으로 어떤 충격**을 주는지 식별하는 것입니다.

[슬라이드5]

저희는 3기 신도시의 아파트 가격을 예측하기 위해 **Hedonic 회귀 모델**을 기본으로 사용했습니다.
가격 데이터는 로그 변환이후, 다섯번의 교차검증으로 정확도를 높였고, 7퍼센트 내의 예측 오차 범위 내의 성능과 해석력을 확보했습니다.

그다음은 **Launch Trajectory**인데요, 분양 시점을 기준으로 시간축을 맞춘 후, 가격이 입주 후 어떻게 움직이는지를 **시각화**하여, 3기 신도시의 **미래 가격 흐름을 시뮬레이션 했습니다.**

마지막으로는 **금리나 정책 같은 외부 변수의 영향**을 보기 위해 **Difference-in-Differences 기법**도 함께 사용할 계획입니다.

[슬라이드6]

저희는 프로젝트를 통해 **세 가지 실질적인 가치를 제공하기 위해 노력했습니다.**

첫째, **시장 이해를 높이려 합니다.**
공급, 수요, 정책, 교통 같은 요소들을 함께 고려하여 **주택 가격이 형성되는 복잡한 과정을 계량적으로 설명**했습니다.

둘째, **정책 및 수요자 지원.**
지금 진행 중인 3기 신도시의 분양가나 향후 재판매 가치를 **수치적으로 시뮬레이션**해서, 정책 결정자나 수요자 모두에게 **데이터 기반 판단 도구**를 제공합니다.

마지막으로는 **재현성과 확장성**입니다.
분석 전 과정을 **스크립트화하고 GitHub에 업로드**했기 때문에, 누구든 교육이나 산업 현장에서 **그대로 재사용 가능한 사례**로 활용할 수 있도록 하였습니다.

[슬라이드7]

지금부터 데이터를 어떻게 구성했는지를 설명드리겠습니다.

데이터셋 구성을 위한 전체 흐름은 크게 세 단계,
설계 → 정규화 → 분포 안정화로 나뉩니다.

먼저, 설계 단계에서는 다양한 원시 데이터를 단지 ID와 연월 기준으로 다시 정리해서, 하나의 단지가 시간에 따라 어떻게 변화하는지를 볼 수 있도록 **패널 구조로 재구성**했습니다.

다음으로는 준공연도나 용적률같은 단지 고정 정보와 거래가, 금리, 미분양 등 시간에 따라 바뀌는 정보를 **같은 레벨에 정리**하고, 이 구조를 MultiIndex로 고정해서, 나중에 어떤 변수든 쉽게 추가하거나 확장할 수 있게 만들었습니다.

마지막으로 로그 변환과 분포 보정을 통해, Hedonic 회귀나 이벤트 스터디 모델이 **정규성 가정을 위반하지 않도록** 준비했습니다.

[슬라이드8]

거래가격은 원래 분포의 왜도가 4.1로 오른쪽으로 굉장히 치우진 모습을 보였습니다.
그래서 먼저 **로그 선형화**를 진행해서 ln_price라는 새로운 변수를 만들었습니다.
변환 전후 히스토그램을 비교해 보니, 로그를 씌운 후에는 0 이하로 내려가는 값이 없고, 분포도 훨씬 안정적인 걸 확인했습니다.

그런데 단순 로그만으로 충분한지 판단하기 위해, **샤피로-윌크 검정과 박스-콕스 람다 계산**을 같이 진행했어요.
그 결과, **로그와 비슷한 분포**였지만, 람다가 6정도의 값을 보여 ln_price 자체는 **너무 flat한** 상태였기 때문에, 단순 로그로는 부족하고 **"다시 조정이 필요하다"는 결론**을 내렸습니다.

[슬라이드9]

그래서 모델 학습에 최적화된 데이터를 만들기 위해 **두 가지 작업**을 추가로 진행했습니다.

첫 번째는 **요존슨 변환**입니다.
기존 Box–Cox는 양수만 처리할 수 있어서, 음수나 0이 있는 변수는 처리할 수 없었습니다.
그래서 범위가 더 넓고 유연한 Yeo–Johnson을 사용해서, 분포가 왜곡된 수치 변수들을 정규에 가깝게 만들었습니다.

두 번째는 **윈저화**입니다.
전체 값 중 상·하위 1%는 너무 튀는 값이라 모델에 노이즈를 줄 수 있는 outlier라 판단하여 잘라서 처리했습니다.

그 결과, 가격 계열 변수의 왜도는 크게 줄었고,
정규성 검정 결과인 샤피로-윌크 p-값도 정규분포를 충분히 만족하는 수준이 되었습니다.

[슬라이드10]

지금까지 소개한 전처리 체인은 단순히 데이터를 정리하는 것을 넘어서,
모델 분석의 기반이 되는 **범용 입력 포맷을 제공**하는 데 목적이 있었습니다.

첫째, 로그 선형화와 요존슨 변환, 윈저화를 통해
Hedonic 회귀 모델의 정규성 가정 위반을 최소화할 수 있었고,
이는 분석 결과의 신뢰도를 높이는 데 큰 도움이 됐습니다.

둘째, 이벤트 스터디에서 사용한 **타우-스케일 곡선의 잔차도 안정화**돼서,
시간축 기반 분석이 훨씬 부드럽고 해석 가능한 구조로 바뀌었습니다.

마지막으로 이 데이터셋은 **시계열이나 머신러닝 모델에도 바로 넣을 수 있는 입력 구조**이기 때문에,
향후 ML 실험이나 정책 시뮬레이션에도 그대로 활용할 수 있습니다.

즉, 이 체인은 단순 전처리가 아니라 **분석 확장성과 견고함을 동시에 확보하는 핵심 인프라**입니다.