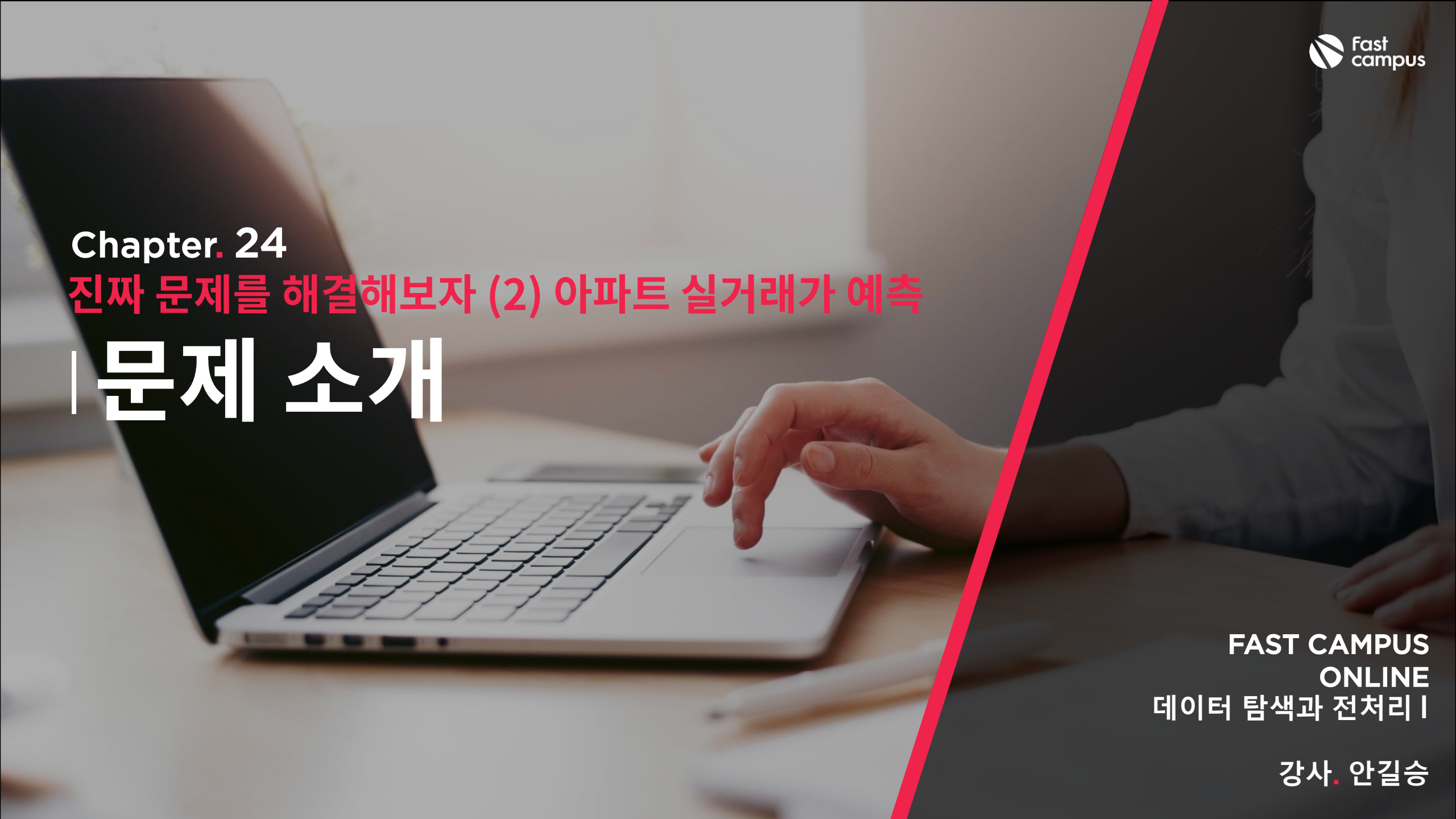


Chapter. 24

진짜 문제를 해결해보자 (2) 아파트 실거래가 예측

문제 소개

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 대회 소개

- 출처: 아파트 실거래가 예측
  - 문제 제공자: 직방 (데이콘)
  - 부동산 빅데이터와 AI를 이용하여 실거래가를 예측 분석
  - https://dacon.io/competitions/official/21265/overview/
- 문제 개요: 서울/부산 지역의 아파트 관련 정보를 바탕으로 실거래가를 예측

I 사용 데이터

- train.csv: 모델 학습용 데이터 (데이터가 커서 샘플링된 데이터 제공) / test.csv: 모델 평가용 데이터
  - apartment_id: 아파트 ID
  - city, dong, jibun, addr_kr: 아파트 주소 관련 변수 (시, 동, 지번, 주소)
  - apt: 아파트 단지 이름
  - exclusive_use_area: 전용면적
  - year_of_completion: 설립 일자
  - transaction_year_month, transaction_date: 거래 년월 및 날짜
  - floor: 층
  - transaction_real_price: 실거래가 (라벨, train.csv에만 존재)

I 사용 데이터 (계속)

- park.csv: 서울/부산 지역의 공원에 대한 정보
  - apartment_id: 아파트 ID
  - city, gu, dong: 아파트 주소 관련 변수 (시, 구, 동)
  - park_name: 공원 이름
  - park_type, park_area: 공원 유형, 공원 면적
  - park_exercise_facility ~ park_facility_other: 공원에 있는 시설
  - park_open_year: 공원 개장년도
  - reference_date: 데이터 기록 일자

I 사용 데이터 (계속)

- day_care_center.csv: 서울/부산 지역의 어린이집에 대한 정보
  - city, gu: 어린이집 주소 관련 변수 (시, 구)
  - day_care_name: 어린이집 이름
  - day_care_type: 어린이집 종류
  - day_care_baby_num ~ CCTV_num: 어린이집의 시설 정보
  - reference_date: 데이터 기록 일자

I 참조 데이터

- 참조 데이터는 대회 문제 해결을 위해, 강사가 직접 수집한 데이터이며, 어떠한 정제도 하지 않았음
- 한국행정구역분류.xlsx
  - ➢ day_care_center에는 동 정보가 없고, train.csv에는 구 정보가 없기 때문에 수집하였음

Chapter. 24

진짜 문제를 해결해보자 (2) 아파트 실거래가 예측

변수 변환 및 부착

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

l 구 변수 부착하기

- 가지고 있는 데이터는 시와 동만 있는데, 다른 데이터에 구만 있어서 **병합이 어려울 것**이라 예상됨
- 따라서 가지고 있는 데이터에 구 변수를 부착해야 함

1. 한국행정구역분류.xlsx 데이터 불러오기 (시트명과 헤더 설정 필요): ref_df
2. 시와 동을 기준으로 데이터를 병합하는 방식으로 구 변수를 부착

| 변수 부착시에 자주 발생하는 이슈 및 해결 방안

- 데이터에 있는 키의 상태 공간이 참조 데이터에 있는 키의 상태 공간에 완전히 포함되지 않아서, 병합을 했을 때 **데이터 크기가 줄어드는 경우**가 존재

- 따라서 키 변수의 상태 공간이 겹치는지를 **np.isin 함수**를 사용하여 확인해야 함

I 불필요한 변수 제거

- 사용하지 않는 변수인 transaction_id와 addr_kr 변수는 미리 삭제하여, 메모리 부담을 줄임
- 모델에 직접 사용되진 않는 다른 변수(apartment_id 등)는 조건부 통계를 사용하여 부착할 예정

| 시간 관련 변수 추출 및 변환

- 아파트가 건축된지 얼마나 되었는지를 나타내는 변수 age를 브로드캐스팅 개념을 활용하여 생성
- transaction_year_month 변수의 타입을 string으로 바꾼 후, str accessor를 사용하여 연도를 추출

l 범주 변수 구간화: floor 변수

- 아파트의 층을 나타내는 floor 변수는 이론적으로는 연속형 변수지만, 범주형 변수로 간주하는 것이 적절함 (2층과 3층 아파트의 가격 차이가 있을까?)

- 다만 범주형 변수라고 하기에는 상태 공간의 크기가 커서 그룹화를 해야 함

- 층에 따른 실거래가의 평균을 나타내는 그래프를 통해 그룹 기준을 확인
  - ➢ Tip 1. 상태 공간의 크기가 작으면 박스 플롯을 그리는 것이 더 바람직함
  - ➢ Tip 2. 정교한 차이를 검정하고 싶으면, 일원분산분석의 사후 검정을 통해 군집화를 하는 것이 좋음

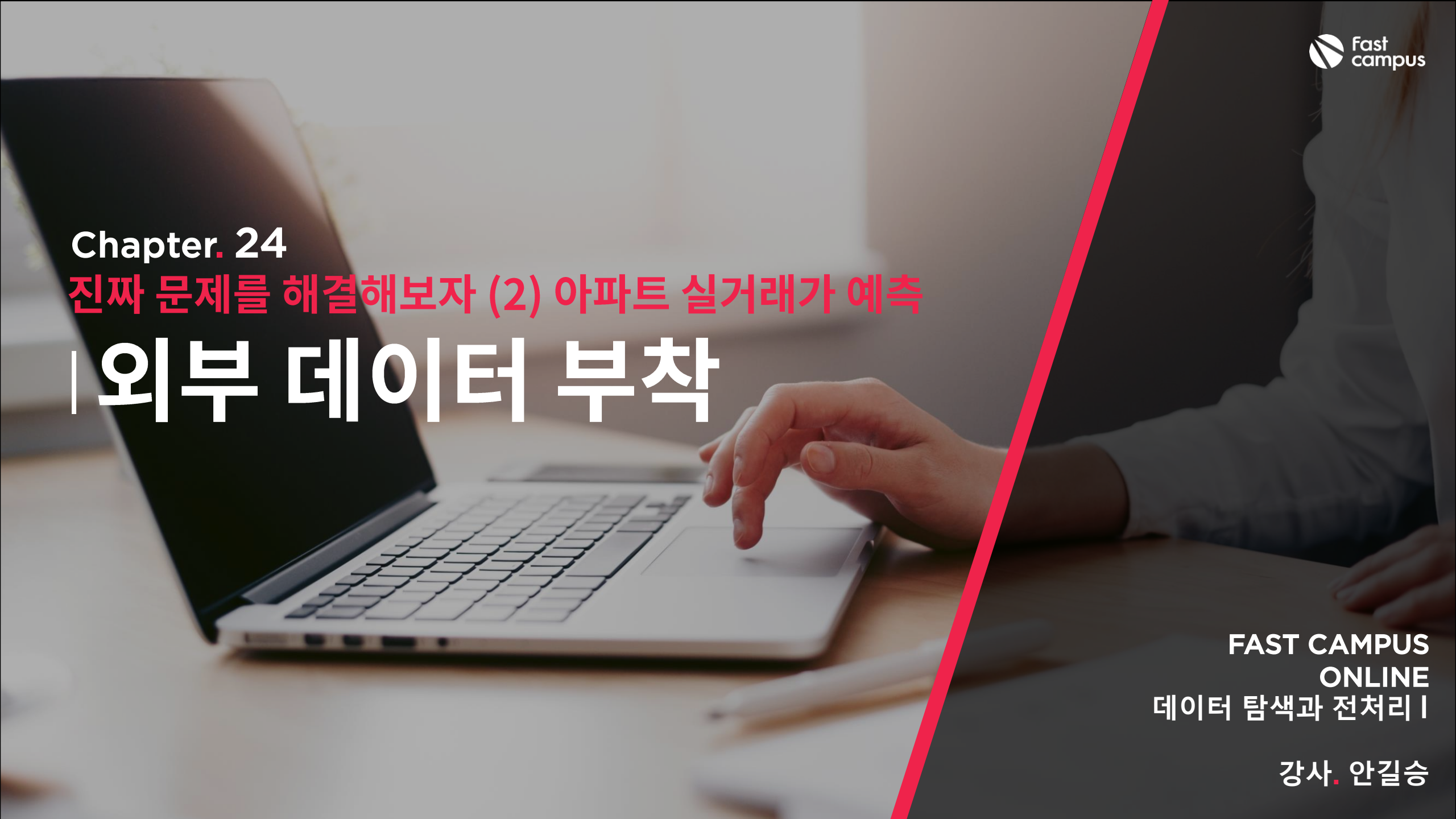
- 그룹 기준을 확인한 뒤, 변수를 구간화하여 각 구간에 속하는 데이터에 대해, 층에 따른 실거래가의 박스플롯을 그려서 군집을 세분화함 (따라서 이전 단계에서 그룹화를 할 때, 여유 있게 나누는 것이 좋음)

I 시세 변수 추가

- groupby를 이용하여 city 변수와 시군구 변수에 따른 transaction_real_price의 평균을 구한 뒤, 이를 데이터에 부착하는 방식으로 **구별_전체_평균_시세 변수**를 추가함

- city, 시군구, transaction_year에 따른 transaction_real_price의 평균과 개수(count)를 구하여, 이를 데이터에 부착하는 방식으로 **구별_작년_평균_시세**와 **구별_작년_거래량 변수**를 추가함

- apartment_id에 따른 transaction_real_price의 평균을 구한 뒤, 이를 데이터에 부착하는 방식으로 **아파트별_평균가격 변수**를 추가함



Chapter. 24

진짜 문제를 해결해보자 (2) 아파트 실거래가 예측

외부 데이터 부착

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

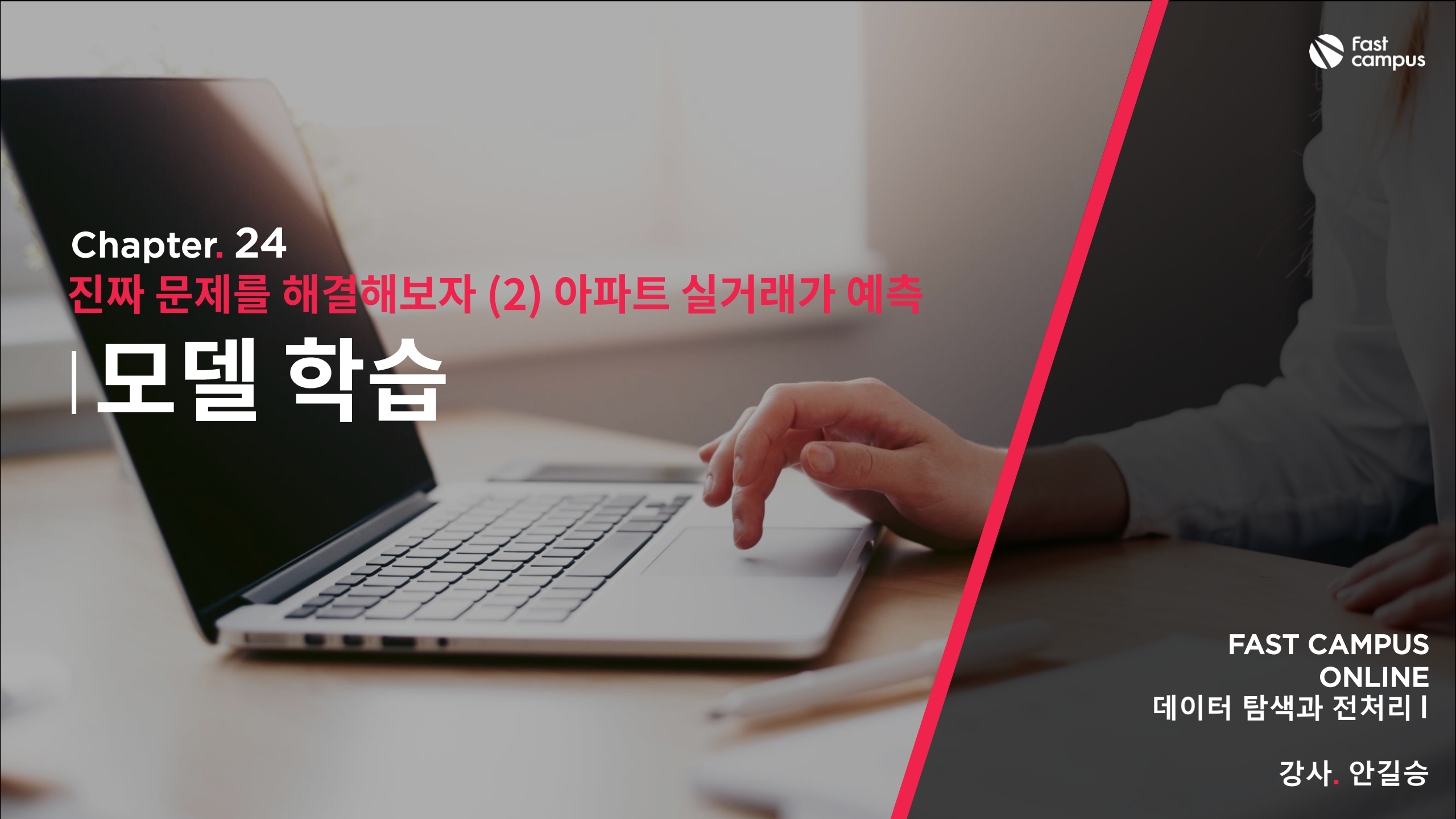
강사. 안길승

I 공원 데이터 추가

- 공원 데이터를 park_df라고 불러옴
- park_df의 park_exercise_facility를 확인해보니, 너무 다양한 종류의 값이 있어 더미화를 하기에는 무리가 있음. 따라서 해당 변수를 결측인지 아닌지만 나타내는 변수로 변환
- 동별로 유형에 따른 공원 수를 계산한 뒤, 데이터에 부착함

I 어린이집 데이터 추가

- 아래 두 개의 가설을 바탕으로, 구 및 유형별 어린이집 수와 케어 가능한 아이 수의 합계를 계산하여 데이터에 부착함
  - ➢ “같은 어린이집이어도 종류가 다르면 아파트 가격에 다르게 영향을 줄 것이다”
  - ➢ “아이가 있는 부모라면, 어린이집 수와 케어 가능한 아이의 수 등만 보고 아파트 구매를 결정하지, 각 어린이집에 CCTV 개수가 몇개인지 등까진 파악하지 않을 것이다”



Chapter. 24

진짜 문제를 해결해보자 (2) 아파트 실거래가 예측

모델 학습

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 데이터 분리

- 라벨 변수에는 transaction_real_price를 할당
- 특징 변수에는 라벨을 포함하여 불필요한 변수를 제거하여 정의
- 또한, 학습 데이터와 평가 데이터를 분할하였으며, 학습 데이터의 크기가 (27012, 22)임을 확인
- 데이터 크기와 변수 타입을 고려하여 트리 계열의 모델을 사용하기로 결정함

| 더미화

- 샘플 대비 특징이 많지 않고, 범주형 변수의 개수도 많지 않아 더미화를 하더라도 큰 문제가 없다고 판단
- 따라서 floor_level 변수에 대한 더미화를 수행 (단, 트리 계열의 모델을 쓰므로 모든 범주 값을 사용)

I 결측 대체

- 원 데이터에는 결측이 없으나, 과거 거래와 관련된 변수를 부착하는 과정에서 과거가 없는 데이터에 대한 결측이 생성되었음

- 또한, 이들 결측이 발생한 변수는 모두 연속형이므로 평균을 기반으로 대체함

I 파라미터 튜닝

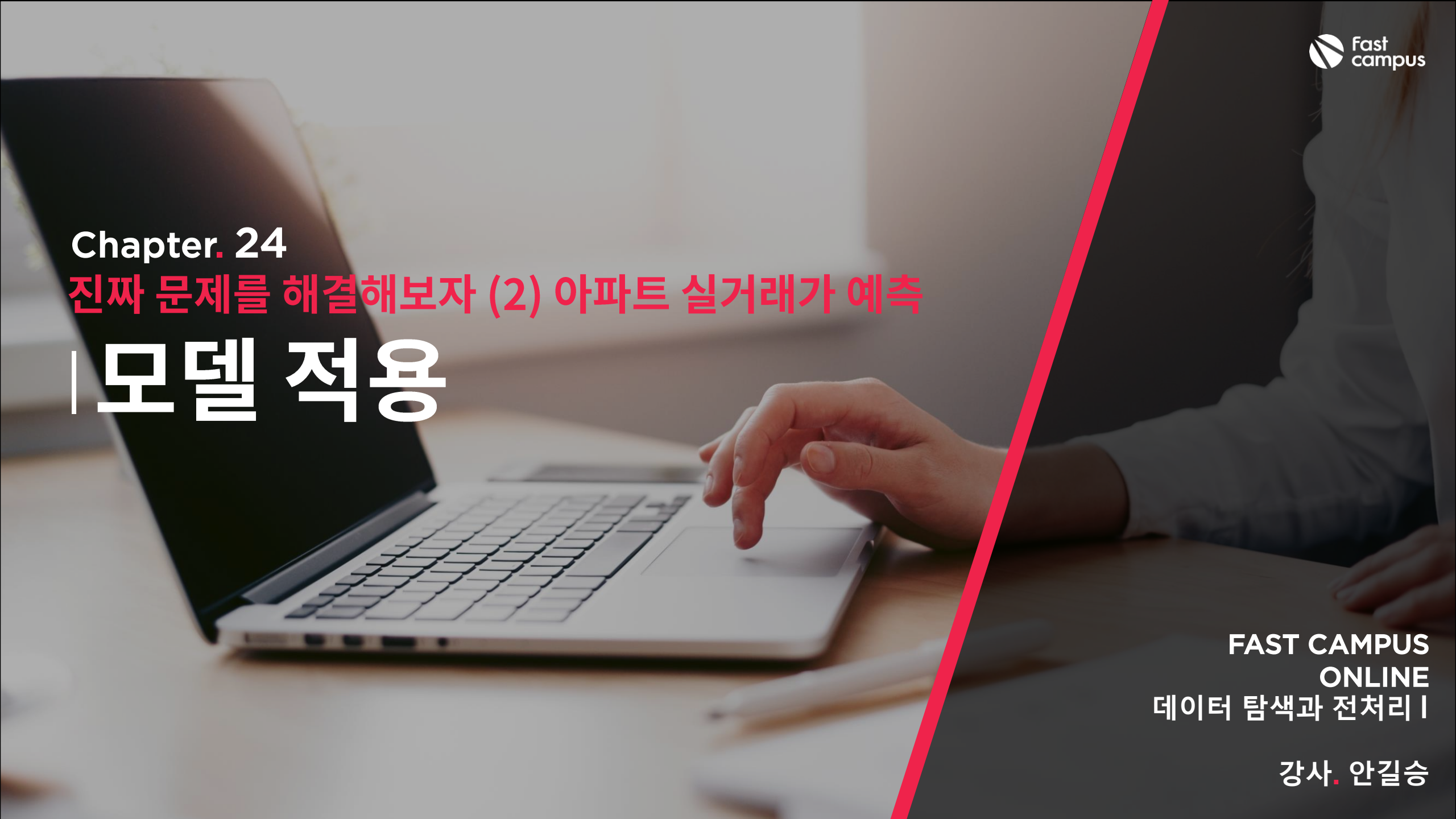
- Random Forest, XGBoost, LightGBM 및 특징 선택에 대한 파라미터 그리드를 다음과 같이 생성

구분	파라미터
Random Forest	➢ max_depth: [3, 4, 5] ➢ n_estimators: [100, 200]
XGBoost	➢ max_depth: [3, 4, 5] ➢ n_estimators: [100, 200] ➢ learning_rate: [0.05, 0.1, 0.2]
LightGBM	
특징 선택	➢ 특징 개수: [20, 15, 10, 5] ➢ 특징 선택 기준: 상호 정보량

- 모든 조합에 대한 파라미터를 그리드 서치 방식으로 튜닝함

I 모델 재학습

- 학습 데이터와 평가 데이터를 pandas의 concat 함수를 사용하여 이어 붙이는 방식으로 데이터를 병합
- 병합한 데이터를 사용하여 모델을 재학습함



Chapter. 24

진짜 문제를 해결해보자 (2) 아파트 실거래가 예측

모델 적용

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

| 파이프라인 구축

- 새로운 데이터에 대한 예측을 수행하기 위해, 하나의 함수 형태로 파이프라인을 구축함
- 또한 파이프라인 사용에 필요한 모든 요소를 pickle을 사용하여 저장하고 불러옴
- 파이프라인을 사용하여 새로운 데이터를 예측함