**팀 프로젝트**

**프로세스 정리**

문제해결을 위한 데이터 분석 시각화 – A

최승호

2071012 이연준

2024. 4. 3.

중간고사대체과제 프로세스 정리

**목차**

1. 프로젝트 주제 설명

2. 프로세스

2.1 데이터 분석 목적 및 목표

2.2 데이터 수집

2.3 데이터 전처리(데이터 정규화)

2.4 데이터 분석

2.5 데이터 활용 및 결과 도출

**1. 프로젝트 주제 설명**

1인 가구 및 자가 소유자가 늘어나고 있는 집값의 문제가 화두로 제기되고 있다. 이에 따라 우리는 집값이 상승하는 원인 및 근거에 집중을 해보기로 하였다. 집을 선정하고 구매할 때 제일 중요하게 생각하는 것 중 하나가 교통편, 즉 집이 역세권에 분포해 있는지라고 생각되었다. 이에 따라 지하철역, 버스정류장 등을 활용하여 역세권에 분포하는 지역을 알아보고, 그 지역에 있는 집들의 가격과 가격상승 폭 및 최적의 주거지라고 평가받는 곳이 실제로 그 근거가 있는지를 확인하려한다. 이 데이터들을 분석하여 위도 경도를 기반으로 시각화할 예정이며, 더 나아가 사용자의 데이터를 기반으로 최적의 주거지를 추천해 주는 시스템을 구축하고자 한다.

**2. 프로세스**

데이터 분석 프로세스는 총 다섯 단계의 프로세스를 사용할 것이다.

데이터 분석 목적 및 목표 – 데이터 수집 – 데이터 정규화(전처리) – 데이터 분석 – 결과 도출(시각화 및 탐색)

위와 같은 5단계를 진행해 프로젝트를 진행하려 하며 세부 과정에 대한 설명은 아래와 같다.

**2.1 데이터 분석 목적 및 목표**

역세권 지역의 집값 상승 및 인구 밀집을 문제점으로 인식하여, 버스정류장 데이터와 지하철역 데이터를 기반으로 집값과 여러 데이터를 분석하여 집값의 상승 및 하강을 예측하고, 더 나아가서는 사용자의 데이터를 기반으로 최적의 주거지를 추천해 주는 시스템을 구축하고자 한다.

**2.2 데이터 수집**

정부에서 운영하는 공공데이터포털(버스정류장 및 지하철역 데이터 수집 – csv파일 형태)을 사용하여 데이터 수집 과정을 줄여 수집된 데이터를 활용할 예정이다. 또한 직방, 다방 등 부동산 관련 웹사이트들을 선정해 웹 크롤링을 사용하여 집값에 대한 데이터를 수집할 예정이다.

**2.3 데이터 전처리 (데이터 정규화)**

버스정류장, 지하철역 데이터에서 중복된 버스정류장 및 지하철역을 제거한다. 이에 따른 분석 목표에 해당하지 않는 불필요한 데이터셋(버스정류장 행정관리 도시명, 지하철 관리 행정자치기구 등)을 제거한다. 또한 웹 크롤링한 데이터에서도 중복된 값을 제거하고, 여러 웹사이트에서 가격 차이가 발생한다면 더 낮은 가격으로 책정된 집값을 데이터로 선정한다.

**2.4 데이터 분석**

버스정류장 및 지하철역 위도, 경도 정보를 기점으로 지역별 로드맵을 그려 분석을 진행한다. 거주지 가격 값 및 기타 정보 등 거주지 관련 정보들을 분석한다. 이후 거주지 정보를 기반으로 거주지가 밀집한 구역을 색 선명도로 구분하여 시각화 및 맵핑을 진행한다. 마지막으로 정류장 및 지하철역 정보를 거주지 정보와 결합하여 한눈에 인구 밀집 및 가격이 보일 수 있게 시각화를 진행하기 위한 상관관계를 생성하여 두 정보의 결합을 진행한다.

**2.5 결과 도출**

분석한 데이터를 기반으로 지도위에 시각화를 진행한다. 지하철역 및 버스정류장 정보와 거주지 정보들을 지도위에 올리고, 이에 따른 집값과 최적의 거주지를 선정하여 시각화로 위치와 정보를 보여준다. 더 나아가 데이터 시각화가 안정화되었다면 원하는 거주지를 탐색할 수 있게 하고, 사용자의 데이터(직장 위치, 주요 활동지, 선호하는 위치 등)를 기재하면 기존 데이터에 근거하여 실질적인 최적의 거주지를 선정하는 추천시스템을 구현한다.