[빅데이터 분석 과정]

Project Guide

강사 이성구

● R 기반 데이터 분석 프로젝트 진행 절차

1. 제안
2. 수집
3. 탐색
4. 전처리
5. 분석
6. 시각화
7. 정확성 검정

1.제안

1)주제 선정

- 공공데이터 기반의 프로젝트를 위한 참신하고 쓸만한 주제 선정

2)타당성 분석

- 우리는 왜 이 주제를 선정하였는가?
- 선정한 주제는 관련 업무를 위해 활용될 수 있는가?
- 선정한 주제는 단순한 데이터 요약이지는 않은가?
- 선정한 주제를 구현할 수 있는 충분한 기술을 확보하고 있는가?
- 선정한 주제를 구현하기 위한 주요 위험 요소는 무엇인가?

3)핵심 속성 및 대상 선정

- 선정한 주제를 성공적으로 구현할 수 있는 핵심 속성 선별
- 분석 결과 활용 대상을 확정하고 무엇을 제공하여야 하는지를 결정

4)팀원 역할 선정

- 프로젝트 관리자: 팀원 작업 지정, 스케줄 정의 및 일정 관리
- 문서 담당자: PPT 및 프로젝트 진행을 위한 각 단계 별 산출물의 작성하고 관리
- 데이터 수집 및 관리자: 데이터 수집 및 통합, 데이터 제약사항 분석
- 시각화 담당자: 분석 결과 시각화, 동적 뷰 생성
- 수석 분석자: 목표한 결과를 얻기 위한 분석 책임자

2. 수집

1) 선정한 주제에 맞는 데이터 수집

2) 핵심 속성을 포함 데이터 수집

3) 공공데이터 사이트 활용

- 공공데이터 포털: http://www.data.go.kr/
- 서울 열린 데이터 광장: http://data.seoul.go.kr/
- 한국복지패널: http://koweps.re.kr/
- 기타 관련 자료 검색 및 수집

3. 탐색

- 수집한 데이터의 내용 확인 및 이해
- 관련 속성 선별
- 데이터 분리 및 통합

4. 전처리

- 자료 별 데이터 프레임 도출 및 통합
- 속성명 & 속성 타입 변경
- 파생 속성 추가
- 결측치 및 이상치 제거 및 정제

5. 분석

- 주제의 목적에 맞는 분석 결과를 얻기 위한 세부적이고 단계적인 분석 전략 수립
- 통계적 가설 검정을 통한 각 단계별 적합성 검증
- 단순한 데이터의 요약이 아닌 데이터 간의 상관 관계를 분석하고 결합하여 수집한 데이터에는 존재하지 않는 새로운 분석 결과 도출

6. 시각화

- 다양한 그래프를 활용한 시각화
- 동적인 인터랙티브한 뷰 제공

7. 정확성 검정

- 최종 결과물이 원래의 목적에 맞게 잘 도출되었는지 검정
- 각 단계별 산출물을 순서대로 검증하여 누락되거나 잘못된 부분이 있는지 조원들 간의 크로스 검정

0. 분석 결과 보고서 작성 및 발표 준비

- 각 단계별 작업 내용을 정리하여 PPT로 작성
- 발표자 선정 및 연습

● 프로젝트 일정

- 2020.06.01 ~ 2020.06.05
- 09:00 ~ 17:50

● 발표

- 2020.06.05 14:00 ~
- 조별 발표 시간 약 30분
- 발표 순서는 당일 사다리로 결정

● 평가

- 각 조별 평가
  - ➢ 각 조별로 다른 조의 발표 결과 평가
  - ➢ 각 조별 평가 점수 20점 * 5조 = 100점
- 4가지 항목에 대하여 평가
  - ➢ 제안의 적합성 – 5점
  - ➢ 분석의 명확성 – 5점
  - ➢ 결과의 정확성 – 5점
  - ➢ 발표의 시각성과 전달력 – 5점
- 1등 조 특전: ???