# DAV 5400 최종 프로젝트 가이드라인

***** 최종 프로젝트를 완료하기 위해 3명 이하의 소규모 그룹으로 작업할 수 있습니다 ***.**

이 과정에서는 다양한 데이터 수집, 데이터 관리 및 데이터 조작 방법을 살펴봅니다. 분석 및 머신러닝 작업의 필수 요소는 새로운 과제가 주어졌을 때 특정 유형의 방법론, 알고리즘 또는 소프트웨어 도구를 언제, 어떻게 적용해야 하는지 아는 것입니다. 이 과정의 **최종 프로젝트에서는** 데이터 소스를 직접 선택하고 연구 질문을 정의할 책임이 있습니다. 작업하고자 하는 데이터 원본을 식별한 후에는 최종 프로젝트 작업의 일부로 답하고자 하는 연구 질문을 하나 이상 정의해야 합니다. 선택한 데이터는 **최종 프로젝트** 작업의 틀을 잡기 위해 정의한 하나 이상의 공식적인 연구 질문에 답하는 데 **사용되어야 합니다.** 이 **최종 프로젝트는 코스 최종 성적의 20%를** 차지합니다.

**최종** 프로젝트는 세 개의 개별 결과물로 구성됩니다:

1. 공식적인 최종 프로젝트 제안서입니다;
2. 최종 프로젝트 글 + Python 코드(Jupyter 노트북 형태);
3. 마지막 라이브 세션(모듈 15)에서 여러분의 작품을 "라이브"로 발표합니다.

이러한 결과물에 대한 일정 및 채점 요약은 아래에 나와 있습니다.

## 결과물 일정

| ***결과물*** | ***날짜*** | ***포인트*** |
|---|---|---|
| 제안서 | 초안: 모듈 9; 최종 초안: 모듈 11 | **25** |
| 최종 프로젝트 | 모듈 15 | **125** |
| 최종 프로젝트 프레젠테이션 | 최종 라이브 세션 중(모듈 15) | **50** |

## 최종 프로젝트 체크리스트

최종 프로젝트에 대한 전체 크레딧을 받으려면 아래 체크리스트에 언급된 모든 항목을 충족해야 합니다. **프로젝트 제안서를 작성하기 전에 이 체크리스트를 주의 깊게 읽어주시기 바랍니다**. 이 체크리스트의 제약 조건 내에서 범위를 제한하고 필요한 단순화 가정을 수행하여 제시간에 작업을 완료할 수 있도록 해야 합니다.

☐ 제안서에는 이 분석을 수행하게 된 동기가 설명되어 있습니다.

☐ 제안서는 데이터의 출처를 설명합니다.

☐ 프로젝트에는 인식 가능하고 재현 가능한 "데이터 과학 워크플로"가 있습니다. [예시: 먼저 데이터를 수집하고 탐색한 다음, 필요한 변환과 정리를 수행한 다음, 분석을 수행합니다.

프레젠테이션 작업이 수행됩니다]

☐ 프로젝트에는 최소 *두 가지* 이상의 서로 다른 *유형의* 데이터 소스(예: 다음 중 두 가지 이상)의 데이터가 포함됩니다: (1) 다운로드 가능한 파일(예: CSV 형식), (2) 스크랩된 웹 페이지, (3) 웹 API. *다음 중 하나 이상*

*소스는 스크랩된 웹 페이지 또는 웹 API 여야 합니다. 쉽게 다운로드할 수 있는 CSV 파일에만 의존하는 것은* 최종 프로젝트에 *허용되지 않습니다.*

- ☐ DAV 5400 과제 또는 프로젝트에 제공된 데이터 세트를 사용할 수 *없습니다.* 또한 Python 라이브러리에 포함된 데이터 세트도 사용할 수 *없습니다*(예,

  모든 scikit-learn 데이터 세트). 또한 데이터는 Kaggle.com에서 *가져와서는 안 됩니다.*

- ☐ 프로젝트에는 데이터를 설명 및/또는 검증하는 통계 분석 및 그래픽(예: EDA)이 포함됩니다.
- ☐ 프로젝트에는 하나 이상의 데이터 재구성 작업이 포함됩니다. [예: 와이드에서 롱으로 변환, 열을 날짜 형식으로 변환]
- ☐ 프로젝트에는 데이터에 적합한 데이터 준비 및 기능 엔지니어링 방법의 사용이 포함됩니다.
- ☐ 프로젝트에는 하나 이상의 그룹화 또는 집계가 포함됩니다.
- ☐ 프로젝트에는 결론을 뒷받침하는 그래픽이 하나 이상 포함되어 있습니다.
- ☐ 프로젝트에는 결론을 뒷받침하는 통계 분석이 하나 이상 포함되어 있습니다.
- ☐ 프로젝트에는 수업 시간에 다루지 않은 파이썬 기능(예: 코스워크 중에 발견했거나 연구를 완료하는 데 필요하다고 생각한 기능)이 하나 이상 포함되어 있습니다.
- ☐ 프레젠테이션. 할당된 시간(8~10분) 내에 프레젠테이션이 진행되었나요?
- ☐ 프레젠테이션. 코드 및/또는 데이터에서 직면한 도전 과제와 그 도전 과제에 직면했을 때 무엇을 했는지 (적어도) 하나 이상 보여줬나요? 도전 과제가 없었다면, 과제는 다음과 같습니다.

  분명히 너무 쉬워요!

- ☐ 프레젠테이션. 청중이 프로젝트를 수행하게 된 동기를 명확하게 이해했나요?
- ☐ 프레젠테이션. 청중이 얻은 인사이트, 도달한 결론 또는 "확인"(거부하거나 거부하지 못한...)한 가설 중 하나 이상을 명확하게 이해했습니까?
- ☐ 코드 및 데이터. 제출한 코드와 데이터를 재현 가능하고 자체적으로 포함할 수 있는 곳(가급적이면 GitHub의 Jupyter Notebook)에 전달했나요? 다음을 사용하여 결과를 완전히 재현할 수 있나요?

  무엇을 제공했나요? 코드가 로컬 컴퓨터의 데이터를 참조하는 경우 전체 크레딧을 받을 수 없습니다.

- ☐ 코드 및 데이터. 전달된 모든 코드가 오류 없이 실행되나요?
- ☐ 마감일 관리. 프로젝트 제안서 초안, 프로젝트, 프레젠테이션이 제시간에 제출되었나요? 제시간에 제출해 주세요! 물론 기한보다 앞당겨 제출하는 것도 가능합니다.

## 협업에 관한 정책

최대 3명으로 구성된 팀으로 **최종 프로젝트를** 완료할 수 있습니다. 각 프로젝트 팀원은 제출된 *모든* 프로젝트 작업물 + Python 코드를 이해하고 설명할 수 있어야 합니다. 다른 곳에서 찾은 작업을 기반으로 삼을 수 있지만, 출처를 인정해야 시작부터가 ***아니라*** 실제로 기여한 내용에 따라 성적이 결정된다는 점을 기억하세요.

---

## 제안 가이드라인

**이 프로젝트의 첫 번째 결과물(25점)은 최종 프로젝트 제안서입니다**. 최종 프로젝트 제안서는 공식적인 연구 제안서 문서 형태로 제출됩니다. . 또한, 제안하는 프로젝트가 **최종 프로젝트에** 명시된 모든 요구 사항을 충족하는지 확인해야 합니다.

**체크리스트(**아래 참조). 제안서에는 아래에 설명된 각 섹션이 포함되어야 하며, Jupyter 노트북의 형태로 제출해야 합니다.

## 소개 (5점)

이 섹션에서는 연구 질문 자체를 실제로 설명하지 않고도 답하고자 하는 연구 질문의 근거에 대한 배경 지식을 제공해야 합니다. 예를 들어, 연구 질문이 건강 관련 문제에 초점을 맞추고 있다면 수집할 수 있는 감염률 또는 사망률 통계를 포함하여 매년 지역, 국가 또는 전 세계적으로 해당 문제로 인해 영향을 받는 사람의 수를 간략하게 요약할 수 있습니다. 기본적으로 서론에서는 제안하려는 연구 질문이 왜 관련성이 있고 독자가 관심을 가져야 하는지 독자가 이해할 수 있도록 해야 합니다.

## 연구 질문(6점)

각 연구 질문을 설명하는 간결한 문장을 한 문장으로 작성하세요. 그런 다음 연구 결과가 '실제 세계'에서 어떻게 사용/구현될 수 있는지 한두 문단 정도 설명하세요.

## 사용 데이터(4점)

데이터의 출처를 명확하게 식별하고 해당 출처에서 데이터를 수집하는 데 사용할 방법을 설명합니다(예: "웹 페이지 스크래핑을 통해 이 출처에서 데이터를 수집해야 합니다..." 등).

## 접근 방식 (10점)

수집하는 데이터를 어떻게 관리할 계획인지 설명하세요(예: 어떤 종류의 데이터베이스에 저장할 것인지 등). 연구 질문에 답하기 위해 어떤 유형의 통계 분석을 활용할 계획인지 설명하세요. 연구 질문에 답하는 데 도움이 되는 그래픽을 생성할 계획이 있다면 설명하세요. 독자가 작업을 어떻게 진행할 계획인지 명확하게 이해할 수 있어야 합니다. 이 글은 제안서이므로 독자에게 제안한 프로젝트가 a) 현실적이며, b) 프로젝트에 할당된 시간 내에 실현 가능하다는 것을 설득할 수 있어야 합니다. 또한 프로젝트 작업에 대한 각 팀원의 역할과 책임을 명확하게 설명해야 합니다.

**최종 프로젝트 제안서 주피터 노트북 결과물은 출판물 수 준의  전문가용 문서와 유사해야 하며, 명확하게 표시된 그래픽, 고품질 서식, 명확하게 정의된 섹션 및 하위 섹션 헤더가 포함되어야 하고 맞춤법 및 문법 오류가 없어야 합니다.**

제공된 최종 프로젝트 제안서 캔버스 제출 포털에서 주피터 노트북을 업로드/제출하세요. 노트북은 반드시 '**이름_성_최종프로젝트제안서**'(예: J_Smith_최종프로젝트제안서)와 같은 명명법을 사용하여 저장해야 합니다. ***소규모 그룹은 주피터 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인해야 하며, 각 팀원은 반드시 캔버스 내에서 팀 작업의 사본을 제출해야 합니다.***

---

## 제안서 승인

제안서를 제출하면 제안한 내용이 이 코스의 최종 프로젝트로 허용되는지 여부를 결정하기 위해 제안서의 내용을 검토합니다. 그렇다면 최종 프로젝트 작업을 시작할 수 있도록 조건부로 승인됩니다. 그렇지 않은 경우 프로젝트 승인을 받기 전에 해결해야 할 문제에 대한 자세한 피드백을 받게 됩니다. 필요한 승인을 받기 위해 필요한 만큼 제안서를 다시 제출할 수 있습니다. 조건부 승인을 받으면 최종 프로젝트의 제안서 구성 요소에 대해 가능한 전체 25점을 획득하게 됩니다(모듈 11에 명시된 마감일 이전에 제안서를 제출했다는 가정 하에).

---

**<u>이 프로젝트의 두 번째 결과물(125점)</u>**은 최종 프로젝트 주피터 노트북입니다. 이 노트북에는 **적절한 형식의 마크다운 셀에** 포함된 Python 코드 셀과 설명 내러티브의 조합이 포함되어야 합니다. 노트북에는 (최소한) 다음 섹션이 포함되어야 합니다(각 섹션의 관련 Python 코드 포함):

1) **초록(10점):** 250단어 이하로 문제, 방법론 및 주요 결과를 요약합니다.
2) **소개(10점):** 답변하기로 선택한 연구 질문에 대한 과학적 또는 비즈니스적 동기를 포함하여 프로젝트를 설명합니다. 이 섹션에서는 최종 프로젝트 제안서의 내용을 요약해야 하므로 연구 질문을 설명하고, 작업하기로 선택한 데이터 세트의 출처와 내용을 설명하고, 프로젝트의 요구 사항을 충족하기 위한 접근 방식을 요약해야 합니다.
3) **연구 접근 방식(10점):** EDA, 데이터 준비 및 조사 분석 작업을 포함하여 최종 프로젝트 작업의 모든 측면에 사용한 엔드투엔드 방법론을 설명하고 제시하세요. 서술의 일부로 데이터 관리 전략에 대한 설명을 반드시 포함해야 합니다.
4) **탐색적 데이터 분석(25점)**: 예비 예측 추론을 포함하여 분석에서 도출한 결론을 포함하여 EDA 작업을 설명하고 제시합니다. 이 섹션에는 EDA에 사용된 Python 코드가 포함되어야 합니다.
5) **데이터 준비(15점)**: 데이터 세트에 적용한 기능 엔지니어링 기법을 포함하여 EDA에서 식별한 데이터 무결성 + 사용성 문제를 해결하기 위해 수행한 단계를 설명하고 보여주세요. 이 섹션에는 데이터 준비에 사용된 Python 코드가 포함되어야 합니다.
6) **준비된 데이터 검토(5점)**: 데이터 준비 후 EDA 분석을 설명하고 제시합니다. 이 섹션에는 데이터 준비 작업 중에 조정한 변수에 대해 EDA를 다시 실행하는 데 사용된 Python 코드가 포함되어야 합니다.
7) **조사 분석 및 결과(40점):** 조사 분석 작업을 설명하고, 그 과정에서 사용된 Python 코드를 포함하여 발표합니다. 조사 질문에 대한 답을 제공하고 설명하세요.
8) **결론(10점):** 연구 내용을 요약하고 연구의 결론을 명확하게 서술하세요. 제안서에서 처음 제기한 연구 질문에 답할 수 있었나요? 프로젝트를 위해 완료한 작업의 향후 확장 가능성에 대해 언급하세요.

**Jupyter 노트북의 결과물은 출판물 수준의 전문가용 문서와 유사해야 하며, 명확하게 표시된 그래픽, 고품질 서식, 명확하게 정의된 섹션 및 하위 섹션 헤더가 포함되어야 하고 맞춤법 및 문법 오류가 없어야 합니다. 또한 Python 코드에는 간결한 설명 주석이 포함되어야 합니다.**

제공된 파이널 프로젝트 캔버스 제출 포털에서 주피터 노트북을 업로드/제출합니다. 노트북은 반드시 '**이름_성_파이널프로젝트**'(예: J_Smith_FinalProject)와 같은 명명법을 사용하여 저장해야 합니다. ***소규모 그룹은 Jupyter 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인해야 하며, 각 팀원은 <u>반드시</u> Canvas 내에서 팀 작업의 사본을***

***제출해야 합니다.***

---

**<u>이 프로젝트의 세 번째 결과물</u>(50점)은** 약 8~10분 분량의 라이브 프레젠테이션입니다. 프레젠테이션에는 연구 질문과 작업하기 위해 선택한 데이터에 대한 간략한 개요, EDA 작업, 데이터 준비 + 기능 엔지니어링에 대한 개략적인 설명이 포함되어야 합니다.

과정, 연구 질문에 답하기 위한 접근 방식에 대한 토론, 연구 결과 및 결론에 대한 설명입니다.