**[ 6월 4주차 프로젝트 수행 일지 ]**

**프로젝트 수행 기간 : 6/19 ~ 6/23**

| **프로젝트 주제** | **서울특별시 내 시니어들의 특성에 따른 문화생활 참여 예측 및 소비 분석** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | **김家네** | **프로젝트 팀원** | **김종원, 김지우, 김연진, 김예리** |

**※**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **김종원** | 1. **국민문화예술활동조사 데이터 전처리** 2. **국민문화예술활동조사 시각화 및 분석**  * **‘장애등록여부’, ‘연령’에 따른** * **‘문화행사 참여 시 어려움’, ‘문화예술 관련 많이 지출하는 항목 및 지출을 늘리고 싶은 항목’에 대한 교차분석** * **‘학력’, ‘종사상지위’, ‘가구소득’, ‘거주지’에 따른 ‘문화예술행사 참여 횟수’의 평균 비교(anova)**  1. **머신러닝 모델링**  * **개인의 특성에 따른 ‘문화예술행사 참여 횟수’의 변화 분석** * **로지스틱 회귀분석 응용** | **23-24 머신러닝 모델링**   * **머신러닝 모델링 완료**   **26 소비분석 데이터**   * **소비분석 데이터 전처리** * **소비분석 데이터 & 머신러닝 결과와 함께 분석**   **27 분석 마무리 및 발표자료 완성**   * **소비분석 데이터 & 머신러닝 결과와 함께 분석 마무리** |
| **김지우** | 1. **국민문화예술활동조사 데이터 전처리**  * **변수명 변경 및 통일** * **가공열 생성** * **범주형 데이터 연도별 기준 통일** * **연도별 데이터 병합**  1. **국민문화예술활동조사 시각화 및 분석**  * **EDA를 통해 각 열에 대한 데이터 분포 확인** * **특정 중요 feature들에 대해 워드클라우드 시각화 진행** * **K-Modes 군집화를 통한 엘보우 커브 확인 , 데이터 파악 및 시각화**  1. **머신러닝 모델링**  * **머신러닝 기법 분담 회의 진행** * **랜덤포레스트 분류 진행** * **머신러닝을 위한 전처리 완료** * **하이퍼 파라미터 튜닝** * **모델 성능 향상을 위해 범주 축소 및 데이터 늘리기 진행** | **23-24 머신러닝 모델링**   * **머신러닝 모델링 완료**   **26 소비분석 데이터**   * **소비분석 데이터 전처리** * **소비분석 데이터 & 머신러닝 결과와 함께 분석**   **27 분석 마무리 및 발표자료 완성**   * **소비분석 데이터 & 머신러닝 결과와 함께 분석 마무리** * **발표자료 완성** |
| **김연진** | 1. **국민문화예술활동조사**  * **도맡은 칼럼 기초 분석** * **데이터 전처리**  1. **2018년과 2020년 칼럼 정의**  * **코드값 엑셀 정리** * **코드값이 다른 부분 정리** * **병합 방안 제안**  1. **종속 변수와 연속형 독립 변수 단순 회귀**  * **5개의 연속형 변수 분포도 시각화** * **봉사활동 횟수 칼럼의 이상 분포 확인** * **연속형 독립변수와 종속변수 간의 단순 선형회귀 분석 [로그 변환 전]** * **연속형 독립변수와 종속변수 간의 단순 선형회귀 분석 [로그 변환 후]**  1. **머신러닝 분석**  * **적용 방식 고민**   + **다중/다항 회귀**   + **PCA** * **LightGBM을 활용한 분류 분석 진행 중** | **23-24 머신러닝 모델링**   * **머신러닝 모델링 완료**   **26 소비분석 데이터**   * **소비분석 데이터 전처리** * **소비분석 데이터 & 머신러닝 결과와 함께 분석**   **27 분석 마무리 및 발표자료 완성**   * **소비분석 데이터 & 머신러닝 결과와 함께 분석 마무리** |
| **김예리** | 1. **국민문화예술활동조사 데이터 전처리**  * **기초 분석(시도, 장애등록여부, 문화예술관련 향후 지출을 늘리고 싶은 항목 1순위, 1년이내 축제 참여 의향)** * **column 정의** * **데이터 전처리**  1. **데이터 시각화**  * **각 열에 대한 시각화**   + **의미 있는 feature 선택** * **지역별 관람횟수 시각화** * **연령/성별 참여 시 어려움 비교** * **범주형 feature와 관람횟수 시각화**   + **연령대별 관람횟수**   + **학력수준별 관람횟수**   + **혼인상태별 관람횟수**   + **가구소득별 관람횟수**   + **종사상지위별 관람횟수**  1. **머신러닝 분석 방법**  * **분석 이후 적용 방법 고민**   + **다중/다항 회귀**   + **로지스틱 회귀**   + **XGBoost** | **23-24 머신러닝 모델링**   * **머신러닝 모델링 완료**   **26 소비분석 데이터**   * **소비분석 데이터 전처리** * **소비분석 데이터 & 머신러닝 결과와 함께 분석**   **27 분석 마무리 및 발표자료 완성**   * **소비분석 데이터 & 머신러닝 결과와 함께 분석 마무리** |

| **2. 강사 피드백** | |
| --- | --- |
| **유선준** | 1. **분석 방향 관련 피드백**  * **단순 회귀 분석**   + **회귀 분석에서 너무 영향이 적은 feature의 경우, R-squared값이 음수로 나올 수 있다고 말씀해주심.**   + **왜곡이 너무 심한 데이터는 이상치 제거와 log 변환을 하지 않고 상관관계 또는 회귀 분석을 진행하는게 좋다고 말씀해주심.** * **EDA 및 시각화**   + **시각화 관련 연령대별 문화예술행사 참여 횟수에 대한 그래프를 추가하면 좋을 것 같다고 말씀해주심.**   + **anova 분석 결과와 EDA를 통한 시각화 결과의 결론이 충돌하는 부분은 하나로 통일하는게 좋다고 말씀해주심.** * **카이제곱 검정 및 anova 분석**   + **anova 분석을 한 것에 대해 기본 가정의 종류에 대한 내용을 추가하면 좋을 것 같다고 말씀해주심.**   + **교차 분석 과 anova 분석을 하게 된 이유를 추가하면 좋을 것 같다고 말씀해주심.**     - **가구 소득에 따른 평균 차이의 결과가 충돌** * **전처리**   + **문화예술행사를 직접 관람한 횟수가 나와있는 열과 매체를 이용하여 관람한 횟수가 나와있는 열을 합쳐서 총 관람횟수 가공열을 만들게 된 이유에 대한 설명이 필요하다고 말씀해주심.**   + **가장 많이 지출한 문화예술행사 분야 열에 대해 가중치 적용이 필요할 수 있다고 말씀해주심.** * **워드클라우드 및 군집화**   + **연령대 별로 (시니어) 필터링 후의 워드클라우드를 시각화해보면 좋겠다고 말씀해주심.**   + **연령대별 군집화 결과를 pie chart, 막대 그래프로 시각화해보면 좋을 것 같다고 피드백 주심.**   + **처음 군집화를 진행했을 때 문자형 범주가 들어간 것에 대해 여쭤봤는데, 숫자형 범주로 만들어야지 머신러닝 학습이 가능한 부분에 대해서 피드백 주심.**   + **K-Modes 군집화도 어차피 군집이 만들어진 이후에는 실루엣 계수를 적용해서 평가 지표 확인 및 시각화를 해보면 좋을 것 같다고 말씀해주심.**  1. **발표 방향성 관련 피드백**  * **지금까지 각자 분담해서 분석을 진행한 결과 유의미하지 않은 결과가 나온 것들도 있는데, 이런 부분은 발표에 포함을 시켜야할지 여부에 대해서 여쭤봄**   + **강사님이 꼭 유의미한 feature가 아니더라도 이러이러한 분석도 진행해보았다. 라고 발표할 수 있다고 말씀해주심.** |