

Growth Hackers 16기| DA Quest

목차

1. 유의사항 및 제출 양식 안내
2. Programming
 - A. quest-1 (Pandas & python)
 - B. quest-2 (알고리즘)
3. Data Analysis
 - A. EDA
 - B. 비즈니스 인사이트 도출

[유의사항 및 제출 양식 안내]

※ DA 퀘스트는 [Programming] 영역과 [Data Analysis] 두 영역으로 구성되어 있으며, [Programming] 영역은 Pandas & Python과 알고리즘 문항, [Data Analysis]는 EDA, 비즈니스 인사이트 도출 문항으로 구성되어 있습니다.

※ 코딩이 처음이신 경우, Codecademy의 기초 Python 강좌 혹은 Jump to Python(박응용 著)을 통해 Python 을 학습하신 후 문제를 해결하실 수 있습니다.

(Jump to Python : <https://wikidocs.net/book/1>)

※ 퀘스트 관련 문의사항은 recruit@ghsnu.com 또는 카카오톡 플러스친구 bit.ly/GH16th_ASK로 보내주시기 바랍니다.

※ '지원서'에 안내된 구글 폼(bit.ly/GH16th_INFO)도 꼭 제출해 주세요.

```
김OO-GH16기-DA-지원서.zip
├── 김OO-GH16기-DA-지원서.pdf
├── 김OO-GH16기-DA-퀘스트폴이.pdf
├── Programming/
│   ├── quest-1/
│   │   └── solution.ipynb
│   └── quest-2/
│       └── sol.ipynb
└── DataAnalysis/
    ├── analysis-solution.xlsx
    └── analysis-solution.ipynb
```

※ 최종 Quest 파일은 지원서 파일과 함께 압축하여 단일한 .zip 파일로 제출해주시길 바랍니다. 파일명은 '[이름]- GH16기-DA-지원서.zip'입니다. (e.g. 김철수-GH16기-DA-지원서.zip)

※ 압축 파일은 '지원서 파일', '퀘스트 폴이 파일', '[Programming] 영역 소스코드 파일', '[Data Analysis] 영역 폴이 파일'을 포함해야 합니다. 세부 디렉토리 구조는 위 예시를 참고 바랍니다.

※ 퀘스트 폴이 파일은 [Programming]과 [Data Analysis] 각 영역의 분량 제한을 지켜 하나의 pdf 파일로 변환 후 제출 바랍니다. 분량 제한을 초과할 시 채점에 불이익이 있을 수 있습니다.

※ 퀘스트 폴이 파일의 파일명은 '[이름]-GH16기-DA-퀘스트폴이.pdf'입니다.

(ex. 김철수-GH16기-DA-퀘스트폴이.pdf)

※ [Programming] 영역 폴이 시, 기타 라이브러리 사용 및 보조함수, 타 모듈 작성이 가능합니다. 타 모듈 및 라이브러리 사용 시, solution.ipynb 또는 sol.ipynb 안에서 import 해 주세요.

※ [Data Analysis] 영역 폴이 파일들(.xlsx, .ipynb 등)은 DataAnalysis 폴더에 모두 넣어주세요.

[Programming]

※ 본 영역은 Python 기초 소양을 측정하기 위한 것이므로 Python을 활용하여 해결해주시길 바랍니다.

※ 풀이코드는 .ipynb파일로 제출 부탁드립니다.

※ 프로그래밍 영역에 대한 pdf 답안은 코드 자체가 아닌 코드의 구조와 로직을 설명해주시길 바랍니다. pdf 답안의 분량은 quest-1, quest-2를 합쳐서 2페이지 이내로 제출해 주세요. 제출하시는 소스 코드의 분량 제한은 없습니다.

[quest – 1 / Pandas & Python]

본 문항은 Growth Hackers에서 자주 쓰이는 Pandas에 대한 기초적 이해도 증진을 목적으로 합니다. 제공된 데이터와 pandas 라이브러리를 활용하여, 아래 문항들을 답변해주시길 바랍니다. (solution.ipynb 파일에 작업하여 그대로 제출 부탁드립니다.)

<데이터 설명>

※ 아래의 데이터들은 실제 데이터에 일부 변형을 가한 형태로 실제 데이터와 차이가 있을 수 있습니다.

cheese_details_url_country_datetime.csv

: cheese.com 웹사이트에 존재하는 각 페이지의 url, 지역, 무게 등의 정보를 담은 데이터

칼럼명	칼럼 데이터 타입	칼럼 내용
url	Object	Cheese.com 웹사이트에서 해당 치즈의 페이지
Country	Object	치즈의 생산 국가
Region	Object	치즈의 생산 지역
weight	Int64	치즈의 무게
datetime	Object	해당 웹사이트에 등록된 일자

cheese_details_0_500.csv, cheese_details_500_1187

: cheese.com 웹사이트에 존재하는 치즈의 특성에 대한 데이터

칼럼명	칼럼 데이터 타입	칼럼 내용
-----	-----------	-------

url	Object	Cheese.com 웹사이트에서 해당 치즈의 페이지
Type	Object	치즈의 유형
Fat_content	Object	치즈의 지방 함량
Milk	Object	치즈 생산에 사용된 우유 종류
Calcium_content	Object	치즈의 칼슘 함량
Texture	Object	치즈의 질감
color	Object	치즈의 색깔

<문제>

문제 0

cheese_details_url_country_datetime.csv, cheese_details_0_500.csv, cheese_details_500_1187.csv를 pd.read_csv를 활용해 각각 data_1, data_2, data_3 데이터 프레임으로 내려 받으세요. (데이터 설명에 존재하지 않는 열은 drop 해주세요.)

문제 1 「테이블 합치기」

문제 1-0

data_1의 datetime 열의 데이터 타입을 pd.to_datetime을 활용해 timestamp로 변경하고 해당 datetime 열의 달(month)를 담은 month 열을 추가하세요. 이 때, dt.month를 활용하세요.

(예시)

	datetime	→	month
row 1	2023-01-07		1
row 2	2023-08-12		8
row 3	2023-12-24		12

문제 1-1

data_2와 data_3를 행을 기준으로 합한 problem_1_1_df 데이터 프레임을 생성하세요. 이 때, ignore_index()를 활용하여 새로운 인덱스를 만드세요.

문제 1-2

문제 1-1에서 생성된 problem_1_1_df와 data_1을 url을 기준으로 inner join한 problem_1_2_df 데이터 프레임을 생성하세요.

문제 2 「결측치 처리」

문제 2-1

문제 1-2에서 생성된 problem_1_2_df 데이터 프레임에 대해 isnull(), mean()을 활용하여 problem_1_2_df의 각 열의 이름을 index로 갖고, 값으로 해당 열의 결측치 비율을 갖는 시리즈 객체를 missing_ratio라는 이름으로 생성하세요.

문제 2-2

문제 2-1에서 생성된 missing_ratio를 활용하여 결측치 비율이 60%를 넘는 열의 이름을 담은 list를 생성하세요. 이를 활용하여 problem_1_2_df에서 결측치 비율이 60%를 넘어서는 열을 제거한 problem_2_df 데이터 프레임을 생성하세요.

문제 3 「함수 생성 및 적용」

문제 3-1

input으로 문자열 또는 null이 주어질 때, input이 null인 경우 0, input이 문자열인 경우 해당 문자열을 쉼표(,)를 기준으로 분리하여 리스트를 생성하고 해당 리스트의 길이를 반환하는 count_mixed 함수를 생성하세요. (hint : split 함수를 사용해보세요.)

(input, output 예시)

(input : null / output : 0) (input : cow, goat / output : 2) (input : cow / output : 1)

문제 3-2

문제 2에서 생성된 problem_2_df 데이터 프레임에 count_mixed 함수를 apply를 통해 milk, country열에 적용하고 각각의 값을 mixed, country_mixed라는 열로 저장한 problem_3_df 데이터 프레임을 생성하세요

(예시)

	milk	country	mixed	country_mixed
row 1		Italy	0	1
row 2	Sheep	France, USA	1	2
row 3	Cow, Goat	France	2	1

문제 4

문제 3에서 생성된 problem_3_df 데이터 프레임에서 country_mixed열과 mixed열이 모두 1인 행들만 포함한 problem_4_df 데이터 프레임을 생성하세요.

문제 5

문제 4에서 생성된 problem_4_df 데이터 프레임의 color 열에 대해 yellow가 포함된 행들의 color 열에 yellow를 배정한 problem_5_df 데이터 프레임을 생성하세요.

이때 re모듈의 compile을 활용하세요.

(예시)

	color	→	color
row 1	pale yellow		yellow
row 2	green		green
row 3	yellow		yellow

문제 6

문제 5에서 생성된 `problem_5_df` 데이터 프레임에 `value_counts()`를 활용해 국가명을 인덱스로

데이터 프레임에서 해당 국가에 해당하는 행의 수를 값으로 담은 시리즈 객체를 생성하세요.

그 다음, 해당 시리즈 객체를 `to_dict()`를 활용하여 dictionary형태로 바꾼 `country_dict`를 생성하세요.

이때, `country_dict`에서 value값을 기준으로 내림차순 정렬해주세요.

문제 7

문제 6에서 생성된 `country_dict`에서 value값이 가장 큰 5개의 key를 담은 리스트 `top_country_list`를 생성하세요.

문제 5에서 생성된 `problem_5_df` 데이터 프레임에서 조건 1 혹은 조건 2를 만족하는 행들을 추출하여 `problem_7_df` 데이터 프레임을 생성하세요.

조건 1: color열의 값이 yellow

조건 2: milk열의 값이 cow이며 country열의 값이 `top_country_list`에 포함

문제 8 [groupby 활용]

문제 8-1

문제 7에서 생성된 `problem_7_df`에서 `groupby`를 활용하여 country별로 weight의 평균 및 표준

편차를 구하고 이를 활용해 country, weight_mean, weight_std를 열로 가지는 `problem_8_1_df`

데이터 프레임을 생성하세요. (주의 : 이때 country는 index가 아닌 열로 들어가야 합니다.)

문제 8-2

문제 7에서 생성된 problem_7_df에서 groupby를 활용하여 country별로 가지는 행의 개수를 구하고 이를 country, count를 열로 가지는 problem_8_2_df 데이터 프레임을 생성하세요. (주의 : 이때 country는 index가 아닌 열로 들어가야 합니다.)

[quest – 2 / Algorithm]

그해이는 더운 여름날, 3명의 친구들에게 공평하게 자신이 가지고 있는 물을 최대한 많이 나눠주고 싶습니다. 그해이는 마침 가지고 있던 3개의 병에 물을 각각 나누어 주려고 합니다. 첫번째 병에 물을 붓고 첫번째 병에 담긴 물을 다시 다음 병에 붓는 방식으로 나눠서, 최종적으로 세 병에 담겨 있는 물의 양을 구하고자 합니다.

아래의 설명을 참조하여 그해이가 가지고 있는 3개의 병을 표현한 리스트가 input으로 주어졌을 때, 최종적으로 병들에 담을 수 있는 최대의 물의 양을 파악하는 함수 `your_algorithm()`을 `sol.ipynb`에 작성하세요.

<유의사항>

- 1) 퀘스트 파일에 주어진 Test Dataset을 활용하여 1000개의 input에 대한 정답을 여부를 파악할 수 있습니다.
- 2) 실제 채점에서는 주어진 Test Dataset 이외의 데이터를 활용하여 채점이 이루어집니다.

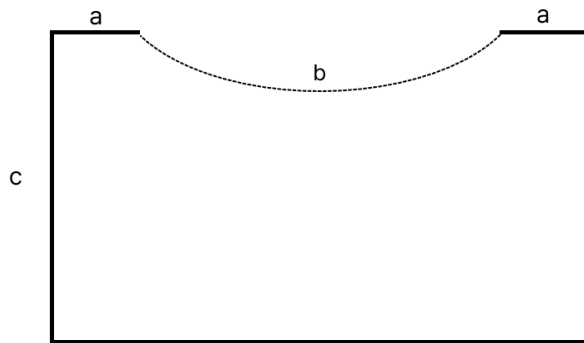
<설명>

- 1) 한 병은 길이가 3이고 1~4사이의 값을 갖는 리스트로 주어집니다.
- 2) input은 3개의 리스트를 담은 1개의 리스트로 주어집니다. 각각의 리스트는 1개의 병을 의미합니다.
- 3) 첫 번째 병에 붓는 물의 양은 항상 40입니다.
- 4) 첫 번째 병에 40의 물을 붓습니다. 첫 번째 병에 물이 차면 이후 병을 90도 기울여 두 번째 병에 물을 붓습니다.
- 5) 두 번째 병에 물이 차면 해당 병을 90도 기울여 세 번째 병에 물을 붓습니다.
- 6) 세 번째 병에 물이 차면 해당 병을 90도 기울여 바닥에 물을 붓습니다.
- 7) 주어진 3개의 병의 순서를 마음대로 변경할 수 있다고 할 때 4,5,6의 과정을 거친 후 3개의 병에 담길 수 있는 최대의 물의 양을 구하세요.

<그림 예시>

예시 1) 리스트에 따른 병의 모양

[a, b, c]



예시 2) 90도 기울인 경우 병 속에 남은 물의 양



예시 3) 4,5,6의 과정 설명

Input = [[3,2,3], [1,1,3], [1,2,4]]

병의 순서 : [3,2,3] -> [1,2,4] -> [1,1,3]

- 첫 번째 병에 40의 물을 부으면 24의 물이 병에 담깁니다.
- 첫 번째 병을 90도 기울이면 15의 물이 두 번째 병으로 쏟아지고 9의 물이 첫 번째 병에 남습니다.
- 두 번째 병에 15의 물이 담깁니다.
- 두 번째 병을 90도 기울이면 11의 물이 세 번째 병으로 쏟아지고 4의 물이 두 번째 병에 남습니다.
- 세 번째 병에 9의 물이 담깁니다.
- 세 번째 병을 90도 기울이면 6의 물이 바닥에 쏟아지고 3의 물이 세 번째 병에 남습니다.
- 최종적으로 16의 물이 3개의 병에 남습니다.

[Data Analysis]

※ 본 영역은 지원자의 데이터 분석 능력을 평가하기 위해 출제되었습니다. 분석을 위한 논리를 중심으로 평가하므로, 도구와 방식은 평가의 대상이 되지 않으며, 별다른 제한을 두지 않습니다.

※ 풀이에 사용하시는 .ipynb, .py, .r, .xlsx 등 소스 코드 및 풀이 도구의 분량 제한은 없습니다.

※ 제공된 데이터 셋을 바탕으로 풀이해 주시고, 외부 데이터를 접목하지는 말아 주시기 바랍니다.

※ 답안의 형태는 다음과 같이 총 5 페이지 이내로 제한해주시길 바랍니다.

- EDA 문항 풀이 최대 2 페이지 (시각화 및 코드 포함)

- 비즈니스 인사이트 도출 문항 1-1, 1-2 번 문항 풀이(시각화 및 코드 포함) 도합 최대 2-3 페이지

- 비즈니스 인사이트 도출 2 번 문항 풀이(시각화 및 코드 포함) 최대 2 페이지

※ 본 문제에서 등장하는 맥락, 상황은 오직 데이터 문제 출제의 편의만을 위해 가정한 것으로, 실제와는 전혀 무관한 완전한 허구이며, 그 어떠한 가치판단도 포함하고 있지 않음을 밝힙니다

[GH Delivery 데이터 분석]

GH Delivery는 모회사 GH팡의 신생 로지스틱스 자회사입니다. 현재 GH Delivery는 모회사 GH팡의 직매입 상품의 배송만을 담당하고 있습니다. GH Delivery 기업의 외형 성장을 도모하기 위해 지금까지 내부적으로 수립한 KPI 달성도를 점검하고, 여러가지 신사업 진출을 검토하는 단계에 있습니다. 이러한 상황을 고려해 주어진 데이터 셋을 활용해 다음의 문항에 답하세요.

<유의사항>

1) 데이터 처리, 가공, 조합, 시각화 방식과 관련 프레임워크 사용은 자유입니다. 하지만 제공된 데이터 셋 이외의 외부 데이터는 도입하지 말아 주시기 바랍니다.

2) 단순히 가설을 수립하고 마치는 것이 아니라, 주어진 데이터 셋을 활용해 도출한 인사이트를 토대로 가설을 뒷받침해주세요.

3) 기업 측에 전달하는 전략 문서와 같이 필요사항, 주의사항 등 최대한 구체적으로 작성해주세요.

4) 데이터 셋에 위도, 경도 데이터가 주어졌지만 해당 지역은 가상지역이라고 가정합니다.

<데이터 설명>

GH_Delivery.csv

: 배달시간, 배달 시 교통상황 등의 정보가 포함되어 있습니다.

칼럼명	설명
Order_ID	배달의 ID (데이터 테이블에서 key로 사용됨)
Agent_Age	배달원의 나이, 현재 고용된 배달원들의 나이는 20~60세
Order_Rating	고객이 해당 배달에 준 평점(1~5점 사이의 값)
Store_Latitude	물건을 보내는 가게(배송 출발지)의 위도
Store_Longitude	물건을 보내는 가게(배송 출발지)의 경도
Drop_Latitude	물건을 받는 고객(배송 도착지)의 위도
Drop_Longitude	물건을 받는 고객(배송 도착지)의 경도
Order_Date	주문 일자
Order_Time	주문 시간
Pickup_Time	배달원이 가게에서 배달물품을 픽업한 시간
Weather	날씨
Traffic	교통상황
Vehicle	배달에 이용된 이동수단
Area	배달 지역
Delivery_Time	배달에 걸린 시간
Category	배달 품목의 카테고리

[EDA]

EDA(Exploratory Data Analysis, 탐색적 데이터 분석)는 데이터의 이상치, 결측치 등을 적절히 처리하고, 데이터의 특징을 올바르게 이해하여 인사이트를 찾아내는 작업을 의미합니다. GH Delivery는 배달에 대한 고객 만족도에 관심이 많기에 Order_Rating과 Delivery_Time에 관심이 많습니다.

<문제>

문제 0

주어진 데이터에서 결측치, 이상치 또는 정합성이 맞지 않는 데이터가 있을 수 있습니다. 만약 있다면 적절히 삭제 혹은 대체하고, 해당 처리 방식에 대한 이유를 설명해주세요.

그리고 처리가 완료된 데이터 프레임을 `processed_df`라는 이름으로 저장해주세요.

문제 1

Category에 따른 Delivery Time의 분포를 파악하고자 합니다. 적절한 방식을 활용하여 시각화를 진행하고 도출한 인사이트를 간략하게 설명해주세요.

문제 2

Delivery Time과 Order_Rating의 관계를 파악하고자 합니다. 적절한 방식을 활용해 시각화를 진행하고 도출한 인사이트를 간략하게 설명해주세요.

[비즈니스 인사이트 도출]

이상치, 결측치 및 정합성에 맞지 않는 데이터를 적절히 처리한 processed_df를 활용하여 아래의 문제들을 해결해주세요.

<문제>

문제 1-1

GH Delivery는 현재 '배송고객 만족도 극대화'를 KPI*로 삼고 있습니다. 현재까지의 KPI 달성도를 점검하고 이를 극대화 할 수 있는 전략을 서술하세요. 이 때, 여러가지 변수를 고려한 배달원 인력 배치에 관한 관점을 포함하세요.

** KPI (Key Performance Indicator) : '핵심 성과지표'로, 비즈니스 목표를 얼마나 잘 달성하고 있는지 판단하기 위해 사용하는 척도*

문제 1-2

2021년 고용노동부의 조사에 따르면 배달원의 주요 연령대는 20대와 30대입니다. GH Delivery는 현재 증가 추세에 있는 배달 주문량에 대비하기 위해 배달원 추가 채용을 고려하고 있습니다. 두 연령대 중 한 연령대를 집중적으로 고용하고자 할 때, 현재 GH Delivery의 KPI와 1-1.에서 도출한 인사이트를 토대로 적절한 결정을 내리고 그 근거를 설명해주세요.

문제 2

GH Delivery는 이익 극대화를 위해 직매입 상품 배송 뿐만 아니라 3자 물류 사업 진출*을 고려하고 있습니다. GH Delivery는 현재 랜덤 스토우* 방식을 도입해 물류를 관리하고 있다는 점을 고려하여 해당 신사업의 사업성을 평가하고, 만약 신사업을 진행하게 된다면 어디에 물류센터를 건설하는 것이 좋을지 제안하세요. 단, 물류센터 건설 시 건설할 지역과 보관할 상품의 특성, 배달 시간을 반드시 고려해야 합니다.

*3자 물류 : 택배 업무를 물류전문기업에 외주를 맡기는 것. 여기서 물류전문기업은 GH 딜리버리를 가리킵니다.

*랜덤 스토우(Random Stow) : 상품의 입출고 시점, 주문 빈도, 물품 특성 등을 인공지능이 종합적으로 판단해 상품을 보관할 위치를 알려주는 시스템