Team Exercise 1

Task 1: Apartment Rent Data

This dataset comprises detailed information on apartment rentals, ideal for various machine learning tasks including clustering, classification, and regression. It features a comprehensive set of attributes that capture essential aspects of rental listings, such as:

Identifiers & Location: Includes unique identifiers (id), geographic details (address, cityname, state, latitude, longitude), and the source of the classified listing.
Property Details: Provides information on the apartment's category, title, body, amenities, number of bathrooms, bedrooms, and square_feet (size of the apartment).
Pricing Information: Contains multiple features related to pricing, including price (rental price), price_display (displayed price), price_type (price in USD), and fee.
Additional Features: Indicates whether the apartment has a photo (has_photo), whether pets are allowed (pets_allowed), and other relevant details such as currency and time of listing creation.
The dataset is well-cleaned, ensuring that critical columns like price and square_feet are never empty. This makes it a robust resource for developing predictive models and performing in-depth analyses on rental trends and property characteristics.

① Find out the average rental price for different room types (2 bed + 1 bath, 3 bed 1 bath, etc).

② Find out the top 10 cities with most expensive and cheapest rental price. Use the price normalized by the size (square feet) for the fairness.

③ Using the normalized rental price from above, compare the average price and variance of each state. Estimate the living expenses for housing of different states. Which state is most affordable? which state would you like to live if you need considering jobs and other environment.

④ Calculate the frequency of different room types. Make a frequency table to compare each state. What are the most common types of apartment of different region?

Task 2: **Top 250 Korean Dramas (KDrama) Dataset**

This dataset contains data from the top-ranked 250 Korean Dramas as per the MyDramaList website. The data has been collected and uploaded in the form of a CSV file and can be used to work on various Data Science Projects.

The CSV file has 17 columns and 251 rows containing mostly textual data.

Most of the data were collected from the MyDramaList website (https://mydramalist.com), and the data for the names of Production Companies was collected from Wikipedia (https://www.wikipedia.org). I wasn't sure how to scrape the data at the time, and hence I went all manual; copying and pasting the data using the cursor. (Yes it was very tedious to manually copy and paste the data!)

I was working on a Content-based Recommender System for Korean Dramas and I needed data to work with. The datasets available on Kaggle had up to only 100 k-drama titles. Not only that, but quite a few of the features deemed essential were also missing; Synopsis, Tags, Director's name, Cast names, Production Companies' names, and such data weren't available with the pre-existing datasets.

① Find out top 10 directors, screenwriters, and production companies with highest average rating? Are you familiar with them? Do you have any favorite directors, screenwriters, or production companies in the list? What are the main characteristics or styles of them?

② Find out top 10 directors, screenwriters, and production companies who worked most dramas om the top 250 dataset? Are you familiar with them? Do you have any favorite directors, screenwriters, or production companies in the list? What are the main characteristics or styles of them?

③ Consider the cast. Who is the actor or actress who stars most of the drama in the datasets? Considering the released year, how the famous actor or actress are changed?

④ Compare OTT and cable TV channel, which one more popular and highly rated? Support your argument with the data? Is it changed over time? how did it change from 2003 to 2022?

Task 3: More about "dplyr"

dplyr is a package in the tidyverse group that provides useful functions for handling data frames. There are many useful functions in dplyr that were not covered during class. Please investigate the following functions, briefly explain their functionality, provide example code and results, and discuss when they would be useful. (You have to make your example codes using the data frame from tasks 1, and 2.)

dplyr은 tidyverse 패키지 그룹에 속한 패키지로 데이터 프레임을 다루기 위한 유용한 함수들을 제공한다. 수업시간에 다루지 못한 함수들 중에도 유용한 함수들이 많이 있는데, 다음 함수들에 대해서 조사한 후, 기능에 대해 간략히 설명하고 예제 코드 및 결과를 통해 어떤 경우에 활용하면 좋을지에 대한 의견을 함께 제시하시오. (예제 코드는 task1, 2에 사용된 data frame을 사용해서 만들어야 됩니다.)

4-1 slice() slice_head() slice_tail() slice_min() slice_max() slice_sample()

4-2

```
mutate_all(.tbl, .funs, ...)

mutate_if(.tbl, .predicate, .funs, ...)

mutate_at(.tbl, .vars, .funs, ..., .cols = NULL)
```

4-3

```
summarise_all(.tbl, .funs, ...)

summarise_if(.tbl, .predicate, .funs, ...)

summarise_at(.tbl, .vars, .funs, ..., .cols = NULL)
```

4-4

across() if_any() if_all()

4-5

sample_n() sample_frac()

reference: https://dplyr.tidyverse.org/