Exercise Sheet 1 – Data Science

INF161 Autumn Semester 2021

Exercise 1 (Down the rabbit hole on Prosecco) *For this exercise, you will evaluate a hypothesis of mine, which is that manufacturers of poor wines try to compensate for the poor taste of their wines by choosing long and fancy names. In particular, you will evaluate the hypothesis for Prosecco wines, since Prosecco is considered by many to be a cheaper version of Champagne, which could mean that Prosecco manufacturers have more to prove than those of other wines. You will be working with the* `winemag-data-130k-v2.csv` *dataset.*

Step 1: *Load the* `winemag-data-130k-v2.csv` *file as a* `pandas` *dataframe. Create a new dataframe composed only of the rows corresponding to reviews of Prosecco wines.*
Tip: Use `DataFrame.loc` *to filter rows, and* `DataFrame.reset_index` *to re-index the rows.*

Step 2: *From the Prosecco dataframe you created in step* 1*, create three new dataframes as follows:*

1. *first dataframe is composed of the rows with more than 90 points and should only have the columns "title," "price," and "points.". Sort this dataframe by price in descending order.*
2. *second dataframe is composed of the rows with less than 84 points and should only have the columns "title," "price," and "points.". Sort this dataframe by price in descending order.*
3. *third dataframe is composed of the rows with (greater than or equal to 84 points) and (less than or equal to 90 points) and should only have the columns "title," "price," and "points.". Sort this dataframe by price in ascending order.*

Finally print the number of rows of each of the three dataframes seperately and print the total number of rows of all three dataframes. Check if the sum matches with the total number of rows in the dataframe created in step 1*.*

Tip: You just need several functions for this such as DataFrame.loc, DataFrame.sort values, '&' operator for merging conditions, len() to get the sum of rows.

Step 3: *Consider each of the data frames of step* 2 *and do as follows:*

Add a new column (defined as the number of characters in the title of the corresponding wine) to the dataframe. Finally print the average number of characters for the dataframe.
Tip: Use DataFrame.map to compute title length for each wine. Spoiler: It seems my hypothesis was wrong, and that the data indicates that the opposite is true, i.e., better Prosecco, on average, have longer names :)

Exercise 2 (The ramen king) *For this exercise, you will figure out which country has the best ramen, as judged by reviews in that country.*

Step 1: *Load the* `ramen-ratings.csv` *file as a* `pandas` *dataframe. Next, group the rows of the dataframe by country (using* `DataFrame.groupby`*) and compute the average number of stars for each country. Also compute the 25-th and 75-th quantiles. Save these as columns named as "mean", "q25" and "q75" in a new dataframe indexed by country which is in sorted order(descending order) of mean. Check who has the best ramen(which is basically the top most country)? Since Norway is not part of the data set we have not eliminated the possibility that Norway is the ramen king, although I think we can agree on that the chance of that being true is quite slim.*

Tip: Stars is not given as a float; you need to convert it. Kaggle has a great tutorial on how to use `DataFrame.groupby` *at* `https://www.kaggle.com/residentmario/grouping-and-sorting`.

Step 2: *Since we are anyway looking at this data set we may as well squeeze some more information out of it. Let us next answer the question of which style of ramen is most popular by country. In particular, you should, separately for each country, compute the fraction of reviews of each style of ramen.*

Start by creating a new clean dataframe by loading the `.csv` *file from disk again and proceed as follows:*

- *group the rows by both country and style.*
- *Compute the total number of reviews for each country.*
- *Compute the total number of reviews for each country and style.*
- *Finally, normalize each entry by the total number of reviews from that country.*

Tip: You just need `DataFrame.groupby` *and division and only a few lines of code (around 6). The challenge is to use groups correctly.*

Notice that sum of the entries of column containing normalized values of every country is 1.

Exercise 3 (Restaurants; more items is more good) *For this exercise, you will analyze who orders more items at restaurants, men or women. You will be working with the* `order.csv` *and* `customers.csv` *datasets.*

Step 1: *Start by loading the* `order.csv` *and* `customers.csv` *files into separate dataframes. Now, you need to merge the two dataframes to compute the statistics we want. Note that there is a customer_ID column in both of the dataframes. Use the* `DataFrames.join` *method to merge the two dataframes. Next, select the "item count" and "gender" columns and sort this dataframe by "item count" in ascending order.*
Tip: Kaggle has a good tutorial on how to use join; see `https://www.kaggle.com/residentmario/renaming-and-combining` .

Step 2: *Finally, compute the average number of items separately for men and women. As it turns out, men and women order almost the same number of items.*
Tip: The spelling of male and female is not consistent.