

```

#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Aug 31 09:42:40 2021

@author: edyta
"""

import pandas as pd
import numpy as np

# Exercise 1
vines = pd.read_csv("winemag-data-130k-v2.csv")

# step 1
prosecco = vines.loc[vines.variety == "Prosecco"]
prosecco.reset_index

# step 2
# a
df1 = prosecco.loc[(prosecco['points'] > 90), ['price', 'title', 'points']]

df2 = (prosecco.loc[(prosecco['points'] < 84), ['price', 'title', 'points']])
df2_sorted = np.sort(df2.points)[::-1]

df3 = (prosecco.loc[(prosecco['points'] >= 84) & (
    prosecco['points'] <= 90), ['price', 'title', 'points']])
df3_sorted = np.sort(df3.points)[::-1]

df1_rows = len(df1.index)
df2_rows = len(df2.index)
df3_rows = len(df3.index)
print("Number of rows in df1: ", df1_rows)
print("Number of rows in df2: ", df2_rows)
print("Number of rows in df3: ", df3_rows)

print("Number of filtered dataframes, df1, df2, df3 is equal to number of
rows in prosecco df: ",
    len(prosecco.index) == (df1_rows + df2_rows + df3_rows))

prosecco['title_character_len'] = prosecco['title'].apply(
    lambda x: np.sum([len(y) for y in x]))
print("Average number of characters in wine title (including white spaces): ",
    prosecco['title'].apply(lambda x: np.mean([len(y) for y in x])))

# Exercise 2 (The ramen king)
ramen = pd.read_csv("ramen-ratings.csv")
ramen.drop(index=ramen[ramen['Stars'] == 'Unrated'].index, inplace=True)
ramen['Stars'] = ramen['Stars'].astype(float)
ramen_new_stars = pd.DataFrame(ramen.groupby(['Country'])['Stars'].mean())

ramen_new_stars = ramen_new_stars.rename(columns={"Stars": "mean"})
ramen_new_stars['q25'] = pd.DataFrame(
    ramen.groupby(['Country'])['Stars'].quantile(0.25))
ramen_new_stars['q75'] = pd.DataFrame(
    ramen.groupby(['Country'])['Stars'].quantile(0.75))
ramen_new_stars = pd.DataFrame(ramen_new_stars)
ramen_new_stars = ramen_new_stars.sort_values(by='mean', ascending=False)

print("The best ramen has Brazil")

```

```

ramen.groupby(['Country', 'Style'])
summary_country = ramen.groupby(['Country']).count()
summary_country_style = ramen.groupby(['Country', 'Style']).count()
norm = ramen.groupby(['Country', 'Style']).apply(
    lambda x: x['Review #']/x['Review #'].sum())

```

### # Exercise 3

```

orders = pd.read_csv("orders.csv")
customers = pd.read_csv("customers.csv")

customers = customers[customers.columns[0:2]]
orders = orders[orders.columns[1:3]]

customers['gender'] = (customers['gender'].str.upper())
customers['gender'].str.strip('gender')
customers['gender'].dropna()
orders['item_count'].dropna()

customers.drop(index=customers[customers['gender']
                               == 'nan'].index, inplace=True)

customers.drop(index=customers[customers['gender']
                               == '?????'].index, inplace=True)
customers.drop(
    index=customers[customers['gender'] == ''].index, inplace=True)
customers.drop(index=customers[customers['gender'] == ''].index, inplace=True)

joined = orders.join(customers.set_index(
    'akeed_customer_id'), on='customer_id')
joined.drop(index=joined[joined['gender'] == 'nan'].index, inplace=True)
df = joined
nan_value = float("NaN")

df.replace("", nan_value, inplace=True)
df.dropna(subset=["gender"], inplace=True)
df.dropna(subset=["item_count"], inplace=True)
tricky = df.gender.unique()

df['gender'] = df['gender'].str.replace('FEMALE', 'F')
df['gender'] = df['gender'].str.replace('MALE', 'M')
df['gender'] = df['gender'].apply(lambda x: x.strip())

statistikk = df.groupby(['gender'])['item_count'].mean()

```