Part 1: Data preview

```
#import libraries
import numpy as np
import pandas as pd
import missingno as msno
# msno.matrix(df_ca)
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme(palette=sns.color_palette("ch:s=.25,rot=-.25"))
# sns.set_theme()
```

```
[3] df_ca=pd.read_csv('https://raw.githubusercontent.com/yyannielin/DataFest2021/main/CA/ca.csv')
    df_ca.head()
    #df_ca.shape
```

	DATE	STATUS	QLANG	DEM_GENDER	DEM_AGE	DEM_ABOR	DEM_ABOR_TYPE	DEM_STDNT	DEM_VET	DEM_HEALTH	DEM_LOCATION	DEM_REGION	DEM_POSTAL	DEM_MARITAL	DEM_
0	2017-09-28 18:13:25	3	1	1	18	0	NaN	0	0	0	2	1	B0N	6	
1	2017-10-18 18:46:34	3	1	1	18	0	NaN	1	0	0	1	1	A1B	6	
2	2017-09-29 06:58:50	3	1	1	24	0	NaN	0	0	0	4	1	E5C	6	
3	2017-10-03 17:58:19	3	1	1	16	0	NaN	1	0	0	4	1	E3A	6	
4	2017-10-25 21:09:49	3	1	1	21	0	NaN	1	0	0	2	1	A9A	6	

5 rows × 185 columns

Part 2: Questions and Analysis

1. Education:

does a person's degree of education affect NMU? Among each group, how many different drugs are a person taking with prescription vs. taking as NMU? Are some drugs more widely used than others?

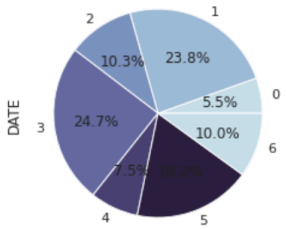
```
# meaningful variables for education:
# *DEM_EDU

# RXDRUGSAFE_HIGH
# RXDRUGSAFE_PAIN
# RXDRUGSAFE_THER

# DRSHOP_NMU
# DRSHOP_SELL

# DAST_1
# DAST_2
# …
# DAST_SUM
```

[9]
```
#DEM_EDU - Key
# Less than high school diploma or its equivalent 1
# High school diploma or a high school equivalency certificate  2
# Trade certificate or diploma  3
# College, CEGEP or other non-university certificate or diploma (other than trades certificates or diplomas)…4
# University certificate or diploma below the bachelor's level  5
# Bachelor's degree (e.g. B.A., B.Sc., LL.B.) 6
# University certificate, diploma, degree above the bachelor's level  7
```

[10]
```
df_ca.groupby('DEM_EDU',as_index=False).count()[['DEM_EDU','DATE']]
```

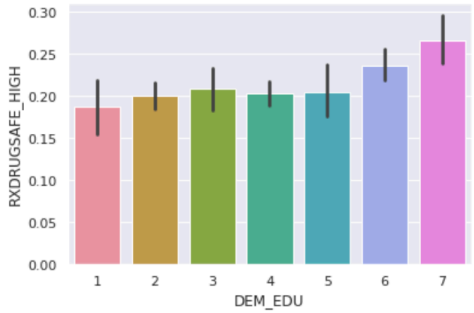
	DEM_EDU	DATE
0	1	549
1	2	2386
2	3	1031
3	4	2471
4	5	750
5	6	1819
6	7	1001

```
num=df_ca.groupby('DEM_EDU',as_index=False).count()['DATE']
num.plot.pie(autopct="%.1f%%");
```



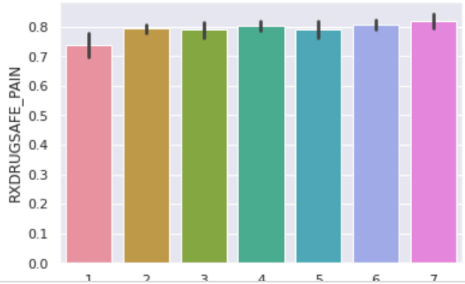

```
[12] #TODO: fit a line to bar chart
     sns.barplot(x= "DEM_EDU", y="RXDRUGSAFE_HIGH",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8eeacf5d90>
```




```
[13] sns.barplot(x= "DEM_EDU", y="RXDRUGSAFE_PAIN",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f41a2a8a210>
```



DataFest2021.ipynb

File Edit View Insert Runtime Tools Help All changes saved

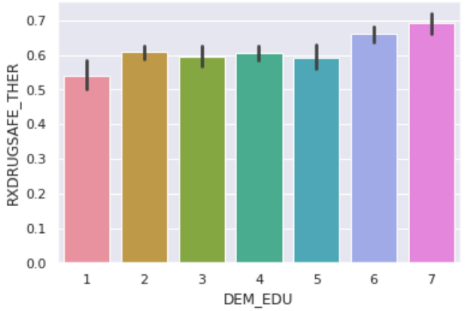
Comment Share A

+ Code + Text

RAM Disk Editing

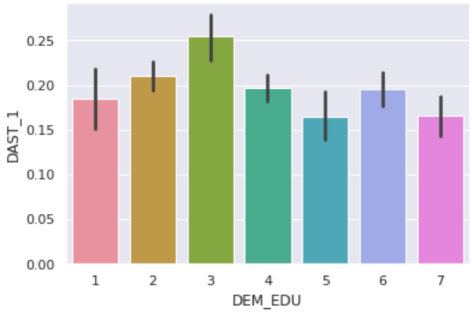
```
[14] sns.barplot(x= "DEM_EDU", y="RXDRUGSAFE_THER",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f41a2a14d90>
```




```
sns.barplot(x= "DEM_EDU", y="DAST_1",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f41a3026750>
```



[17]
```
df_ca.groupby('DEM_EDU').agg({'RXDRUGSAFE_HIGH':'mean','RXDRUGSAFE_PAIN':'mean','RXDRUGSAFE_THER':'mean',
                              'DAST_1':'mean','DAST_2':'mean','DAST_3':'mean','DAST_4':'mean','DAST_5':'mean',
                              'DAST_6':'mean','DAST_7':'mean','DAST_8':'mean','DAST_9':'mean','DAST_10':'mean','DAST_SUM':'mean'})
#TODO: run regression for variables
```

DEM_EDU	RXDRUGSAFE_HIGH	RXDRUGSAFE_PAIN	RXDRUGSAFE_THER	DAST_1	DAST_2	DAST_3	DAST_4	DAST_5	DAST_6	DAST_7	DAST_8	DAST_9	DAST_10	DAST_SUM
1	0.187614	0.735883	0.540984	0.183971	0.027322	0.723133	0.051002	0.109290	0.065574	0.058288	0.041894	0.081967	0.047359	0.943534
2	0.199916	0.792540	0.608550	0.209556	0.039396	0.773261	0.038977	0.092205	0.039396	0.036463	0.034786	0.065801	0.032272	0.815591
3	0.208535	0.789525	0.596508	0.254122	0.055286	0.752667	0.053346	0.110572	0.064016	0.047527	0.058196	0.079534	0.036857	1.006790
4	0.203561	0.801700	0.605828	0.197086	0.047754	0.746257	0.056657	0.089437	0.039255	0.035208	0.052206	0.075273	0.030352	0.876973
5	0.205333	0.789333	0.592000	0.164000	0.080000	0.693333	0.086667	0.070667	0.038667	0.020000	0.076000	0.086667	0.037333	0.966667
6	0.236394	0.806487	0.659703	0.195162	0.033535	0.738868	0.035734	0.076416	0.027488	0.020891	0.028587	0.041781	0.021440	0.742166
7	0.265734	0.818182	0.691309	0.165834	0.027972	0.704296	0.031968	0.077922	0.028971	0.017982	0.024975	0.039960	0.026973	0.738262

[19]
```
# analyze usage of each drug
```

```
# FENT_NMU
data_fent=df_ca[['DEM_EDU','FENT_NMU']].dropna()
mean_fent=data_fent.groupby('DEM_EDU',as_index=False).mean()
# BUP_NMU
data_bup=df_ca[['DEM_EDU','BUP_NMU']].dropna()
mean_bup=data_bup.groupby('DEM_EDU',as_index=False).mean()
# METH_NMU
data_meth=df_ca[['DEM_EDU','METH_NMU']].dropna()
mean_meth=data_meth.groupby('DEM_EDU',as_index=False).mean()
# MORPH_NMU
data_MORPH=df_ca[['DEM_EDU','MORPH_NMU']].dropna()
mean_MORPH=data_MORPH.groupby('DEM_EDU',as_index=False).mean()
# OXY_NMU
data_OXY=df_ca[['DEM_EDU','OXY_NMU']].dropna()
mean_OXY=data_OXY.groupby('DEM_EDU',as_index=False).mean()
# OXYM_NMU
data_OXYM=df_ca[['DEM_EDU','OXYM_NMU']].dropna()
mean_OXYM=data_OXYM.groupby('DEM_EDU',as_index=False).mean()
# TRAM_NMU
data_TRAM=df_ca[['DEM_EDU','TRAM_NMU']].dropna()
mean_TRAM=data_TRAM.groupby('DEM_EDU',as_index=False).mean()
# TAP_NMU
data_TAP=df_ca[['DEM_EDU','TAP_NMU']].dropna()
mean_TAP=data_TAP.groupby('DEM_EDU',as_index=False).mean()
# COD_NMU
data_COD=df_ca[['DEM_EDU','COD_NMU']].dropna()
mean_COD=data_COD.groupby('DEM_EDU',as_index=False).mean()
# COTC_NMU
data_COTC=df_ca[['DEM_EDU','COTC_NMU']].dropna()
mean_COTC=data_COTC.groupby('DEM_EDU',as_index=False).mean()
# HYD_NMU
data_HYD=df_ca[['DEM_EDU','HYD_NMU']].dropna()
mean_HYD=data_HYD.groupby('DEM_EDU',as_index=False).mean()
# HYDM_NMU
data_HYDM=df_ca[['DEM_EDU','HYDM_NMU']].dropna()
mean_HYDM=data_HYDM.groupby('DEM_EDU',as_index=False).mean()
# SUF_NMU
data_SUF=df_ca[['DEM_EDU','SUF_NMU']].dropna()
mean_SUF=data_SUF.groupby('DEM_EDU',as_index=False).mean()
# STIM_NMU
data_STIM=df_ca[['DEM_EDU','STIM_NMU']].dropna()
mean_STIM=data_STIM.groupby('DEM_EDU',as_index=False).mean()
# BENZ_NMU
data_BENZ=df_ca[['DEM_EDU','BENZ_NMU']].dropna()
mean_BENZ=data_BENZ.groupby('DEM_EDU',as_index=False).mean()
# THC_NMU
data_THC=df_ca[['DEM_EDU','THC_NMU']].dropna()
mean_THC=data_THC.groupby('DEM_EDU',as_index=False).mean()


from functools import reduce

dfs = [mean_fent, mean_bup, mean_meth, mean_MORPH, mean_OXY, mean_OXYM, mean_TRAM,
       mean_TAP, mean_COD, mean_COTC, mean_HYD, mean_HYDM, mean_SUF, mean_STIM, mean_BENZ, mean_THC] # list of dataframes

result= reduce(lambda df_left,df_right: pd.merge(df_left, df_right,
                                          left_index=True, right_index=True,
                                          on='DEM_EDU'), dfs)
```

	DEM_EDU	FENT_NMU	BUP_NMU	METH_NMU	MORPH_NMU	OXY_NMU	OXYM_NMU	TRAM_NMU	TAP_NMU	COD_NMU	COTC_NMU	HYD_NMU	HYDM_NMU	SUF_NMU	STIM_NMU	BENZ_NMU	THC_NMU
0	1	0.083333	0.375000	0.111111	0.062500	0.169014	0.200000	0.142857	0.750000	0.262108	0.284900	0.275862	0.129630	0.000000	0.297297	0.103448	0.846154
1	2	0.238806	0.225806	0.227273	0.069620	0.142077	0.394737	0.110345	0.307692	0.245021	0.259505	0.141104	0.133858	0.157895	0.418033	0.099307	0.542373
2	3	0.125000	0.388889	0.280000	0.096899	0.108247	0.147059	0.113924	0.142857	0.229167	0.268229	0.109589	0.108696	0.100000	0.333333	0.119617	0.538462
3	4	0.093023	0.272727	0.204082	0.062136	0.082707	0.233333	0.082474	0.240000	0.237668	0.251682	0.132653	0.084507	0.204545	0.210000	0.082031	0.385965
4	5	0.263158	0.333333	0.363636	0.054217	0.096491	0.280000	0.191176	0.500000	0.291089	0.299010	0.208955	0.118182	0.285714	0.270270	0.088608	0.391304
5	6	0.163636	0.269231	0.250000	0.057402	0.147766	0.265306	0.139860	0.350000	0.202609	0.233308	0.112500	0.087209	0.076923	0.435294	0.093023	0.358974
6	7	0.021277	0.290323	0.222222	0.054945	0.098485	0.173913	0.118280	0.111111	0.165498	0.246844	0.107527	0.076190	0.192308	0.264151	0.105263	0.250000

Conclusion:

Higher education is associated with higher belief in the safety of prescription drugs than illicit drugs DAST_1 measures whether an individual has used drugs other than those required for medical reasons. There is no obvious correlation between an individual's DAST_score and her degree of education. This indicates even though individuals with a higher degree of education understand the safety of prescription drugs, they are not very aware of the potential hazard of consuming prescription drugs without following guidance. Otherwise, higher education would be associated with a lower DAST_1 score.

2. Income

Does the income level of an individual affect NMU? To what extent?

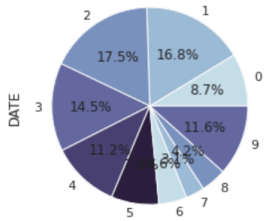
```
[ ] # variable: DEM_INCOME
    #Under $20,000  1
    # Between $20,000 and $39,999 2
    # Between $40,000 and $59,999 3
    # Between $60,000 and $79,999 4
    # Between $80,000 and $99,999 5
    # Between $100,000 and $119,999 6
    # Between $120,000 and $139,999 7
    # Between $140,000 and $159,999 8
    # $160,000 or more  9
    # Prefer not to say 10
```

```
[ ] df_income = df_ca.groupby('DEM_INCOME',as_index=False)
    df_income.count()[['DEM_INCOME','DATE']]
```

	DEM_INCOME	DATE
0	1	866
1	2	1685
2	3	1747
3	4	1450
4	5	1123
5	6	784
6	7	461
7	8	311
8	9	420
9	10	1160

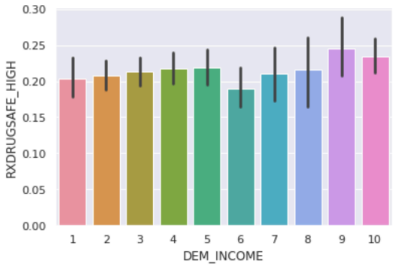
```
[ ] num_income=df_income.count()['DATE']
    num_income.plot.pie(autopct="%.1f%%")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8ee9e88190>
```



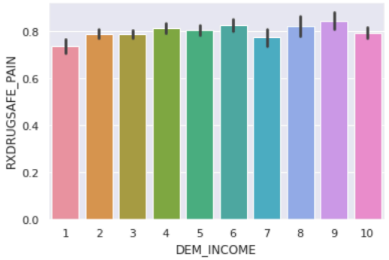
```
[ ] # generate barplots of NMU purpose across income levels
    sns.barplot(x="DEM_INCOME",y="RXDRUGSAFE_HIGH",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f41a293c1d0>
```



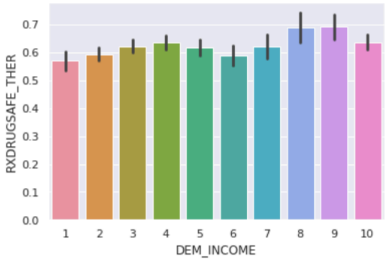

```
[ ] sns.barplot(x="DEM_INCOME",y="RXDRUGSAFE_PAIN",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb24ecaa510>
```




```
[ ] # generate barplots of NMU purpose across income levels
    sns.barplot(x="DEM_INCOME",y="RXDRUGSAFE_THER",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f41a283b250>
```



There is a slight positive correlation between the purpose of NMU and income levels. Relating this to the previous observations on levels of education, it is plausible that there is some degree of collinearity between income and education levels.

3. Gender

```
[ ] # 1: male
    # 2: female
```

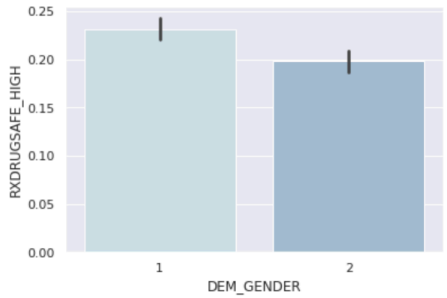
```
[ ] df_gender = df_ca.groupby('DEM_GENDER',as_index=False)
    df_gender.count()[['DEM_GENDER','DATE']]
```

	DEM_GENDER	DATE
0	1	5005
1	2	5002

Fewer women (20%) than men (23%) believe that prescription drugs are safer than illicit drugs for enjoyment.

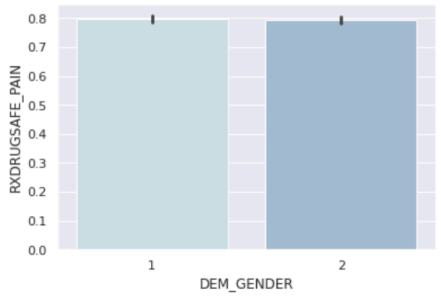
```
[ ] sns.barplot(x= "DEM_GENDER", y="RXDRUGSAFE_HIGH",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f934625ce10>
```

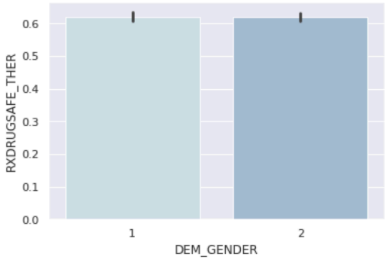



```
[ ] sns.barplot(x= "DEM_GENDER", y="RXDRUGSAFE_PAIN",data=df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9344f28510>
```



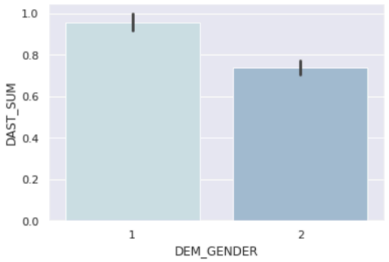
[ ]
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9344e9e050>
```



Does gender have a relationship with attitude and psychological response toward drug use?

```
sns.barplot(x = "DEM_GENDER",y = "DAST_SUM",data = df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f933bd4cb90>
```



Women generally have a more positive attitude and psychological response towards drug use than men, with a mean DAST_SUM of 0.75 compared to 0.95 for men.

This trend is more clearly observed in the side-by-side plot below, where more women than men fall under category 1 in response to drug use, from DAST_CAT 1 representing no negative response reported to DAST_CAT 5 severe negative response.

[ ]
```
sns.countplot(x = "DAST_CAT",hue = "DEM_GENDER",data = df_ca)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f933c83bf50>
```