The Book Café Project

Anupam Misra

Statistics I

PGPDS Spring 2021

Praxis Business school

Contents

"That's the thing about books. They let you travel without moving your feet." – Jhumpa Lahiri, The Namesake [1]

Introduction

While exploring topics for my Statistics project, I came across a number of intriguing ideas. However, I decided to work on something which is personal and would let me exercise my creative muscles.

I belong to a family where someone is always pouring over books, newspapers or magazines. At school, we had a library and a library period dedicated to reading. However, much to the annoyance of my friends and family, I was not a child who was keen on reading.

During college I realized the value of reading books outside the curriculum. To open your mind to black holes and understand the intricacies of human nature, it is important. To quote Mark Twain, "The man who does not read has no advantage over the man who cannot read".

Through discussions with friends about books, I understood the relevance of libraries and book clubs. We as humans are a sum of our past experiences and our uniqueness enables us to generate perspectives not obvious to others. Don't get me wrong, you do not need to dissect the text, analyze and debate about it, but if you choose to do so, it can leave you with a deeper understanding and sometimes, more appreciative of the text.

However, with the advent of the digital age, the popularity of libraries and book clubs seem to be dwindling. But a newer avatar in the form of book cafes is popping up in the metropolitan cities. In this project I have undertaken a study to understand people's reading habits in the broader aim of opening a book café someday. I hope you enjoy reading this project as much as I did while creating it.

Want to head to a book café right now? [1][2][3]

Oxford Bookstore and Cha Bar, Kolkata

Bibliotheque Book Café, Gangtok

Literati Bookshop and Cafe, Calangute, Goa

The Bibliophilia Cafe, Guwahati

Café Turtle, New Delhi

Ivy & Bean, New Delhi

Kitaab Khana, Mumbai

Leaping Windows, Andheri West, Mumbai

Atta Galata, Bengaluru

Illiterati Books and Coffee, Macleodganj

Lehling Bookshop and Coffee House, Leh

Pagdandi Bookstore Cafe, Pune

Book Café Song, Pune

The Coffee Cup, Hyderabad

Books N Brew, Chandigarh

Methodology

The main motive of the study was to find out if people want to read books without buying them.

P = Probability of people wanting to read books without buying them

Sampling

Sampling frame: My contact list - Phonebook and Facebook

Sampling type: Convenience

Sample size(n): 120

Confidence level: 95%

Population proportion(P): 50%

Margin of error: 8.95%

Margin of error = $Z * \sigma/n$

$Z = 1.96$ at 95% confidence

$\sigma = P(1-P)$ for binomial distribution
$P = 0.5$ assumed, i.e., equally likely to swing either way

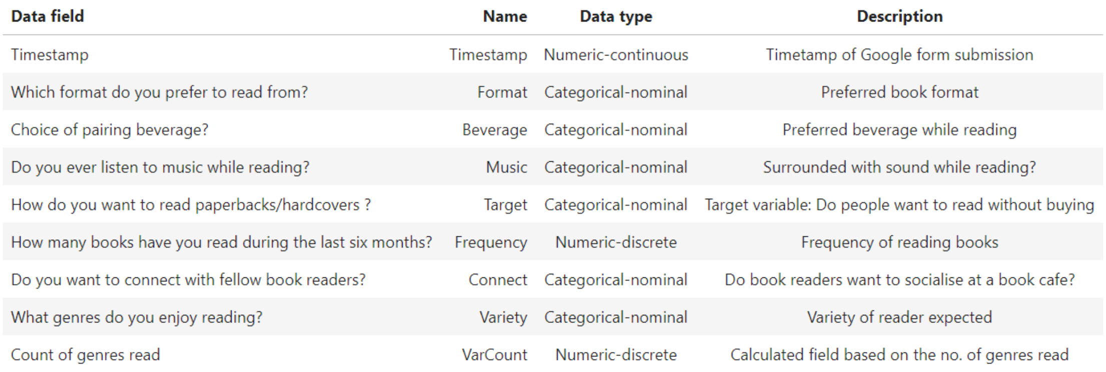
$n = 120$, our sample size

Data collection

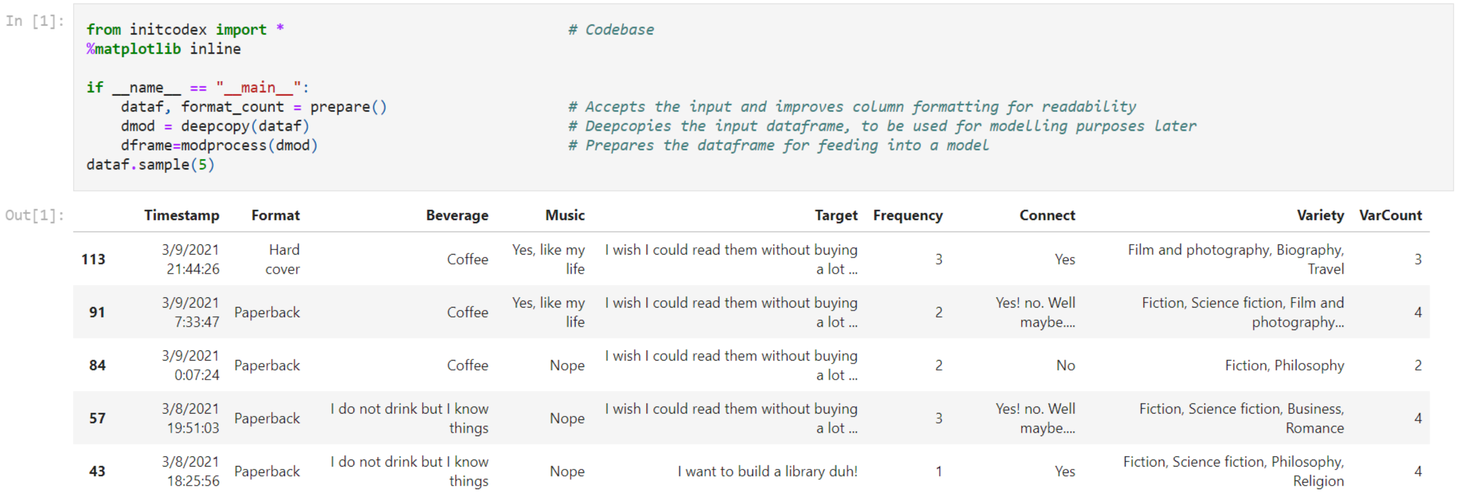
I sent the Google form to all my contacts through different communication channels. I received 120 responses till 15-03-2021. Data was collected anonymously because I wanted the user preferences to be unbiased. Also, there was no need to identify individual preferences. Hence the sentiment of the sample was captured to understand the population sentiment.

The Google form can be accessed here.

Data understanding



Data field	Name	Data type	Description
Timestamp	Timestamp	Numeric-continuous	Timetamp of Google form submission
Which format do you prefer to read from?	Format	Categorical-nominal	Preferred book format
Choice of pairing beverage?	Beverage	Categorical-nominal	Preferred beverage while reading
Do you ever listen to music while reading?	Music	Categorical-nominal	Surrounded with sound while reading?
How do you want to read paperbacks/hardcovers ?	Target	Categorical-nominal	Target variable: Do people want to read without buying
How many books have you read during the last six months?	Frequency	Numeric-discrete	Frequency of reading books
Do you want to connect with fellow book readers?	Connect	Categorical-nominal	Do book readers want to socialise at a book cafe?
What genres do you enjoy reading?	Variety	Categorical-nominal	Variety of reader expected
Count of genres read	VarCount	Numeric-discrete	Calculated field based on the no. of genres read



In [1]:

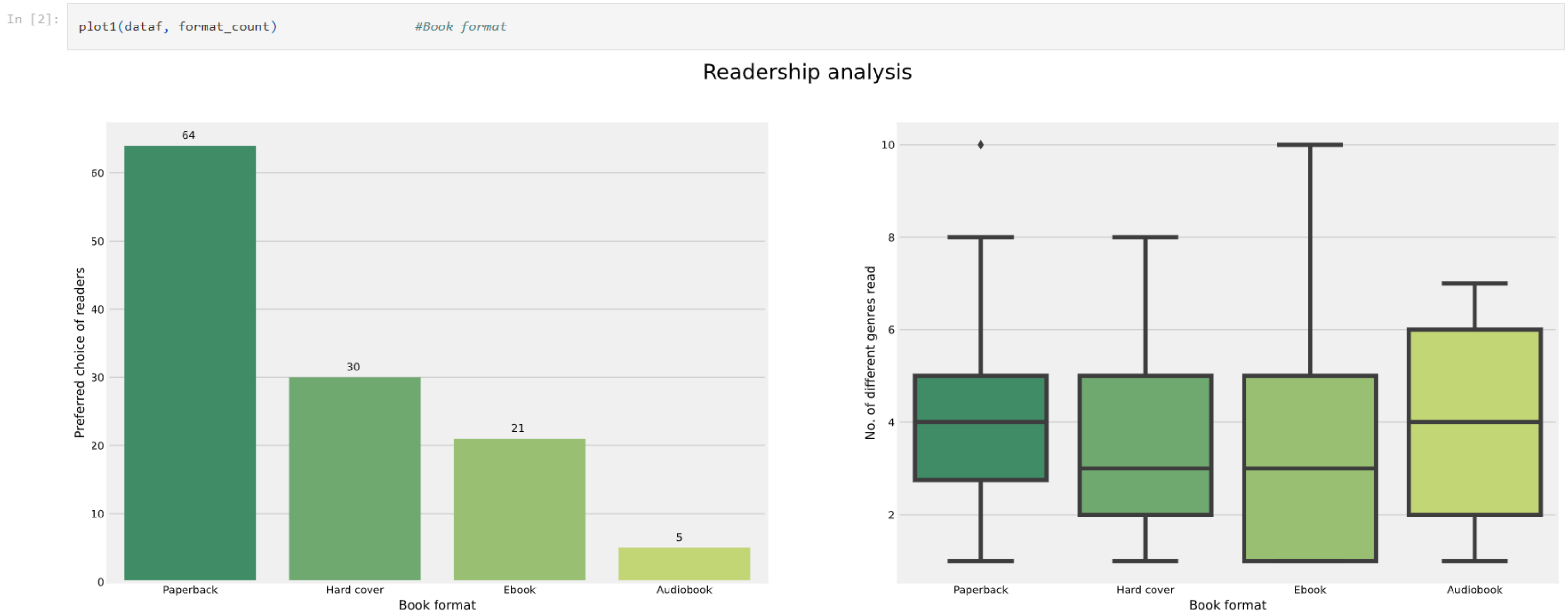
```
from initcodex import *                                       # Codebase
%matplotlib inline

if __name__ == "__main__":
    dataf, format_count = prepare()                           # Accepts the input and improves column formatting for readability
    dmod = deepcopy(dataf)                                    # Deepcopies the input dataframe, to be used for modelling purposes later
    dframe=modprocess(dmod)                                   # Prepares the dataframe for feeding into a model
dataf.sample(5)
```

Out[1]:

	Timestamp	Format	Beverage	Music	Target	Frequency	Connect	Variety	VarCount
113	3/9/2021 21:44:26	Hard cover	Coffee	Yes, like my life	I wish I could read them without buying a lot ...	3	Yes	Film and photography, Biography, Travel	3
91	3/9/2021 7:33:47	Paperback	Coffee	Yes, like my life	I wish I could read them without buying a lot ...	2	Yes! no. Well maybe....	Fiction, Science fiction, Film and photography...	4
84	3/9/2021 0:07:24	Paperback	Coffee	Nope	I wish I could read them without buying a lot ...	2	No	Fiction, Philosophy	2
57	3/8/2021 19:51:03	Paperback	I do not drink but I know things	Nope	I wish I could read them without buying a lot ...	3	Yes! no. Well maybe....	Fiction, Science fiction, Business, Romance	4
43	3/8/2021 18:25:56	Paperback	I do not drink but I know things	Nope	I want to build a library duh!	1	Yes	Fiction, Science fiction, Philosophy, Religion	4

Analysis



```
In [2]: plot1(dataf, format_count)                  #Book format
```

Observations:

1. Maximum readers read from paperbacks or hardcover books. This supports our idea that people would prefer reading from physical books. Hence people might come to our book cafe to read books.

2. Maximum no. of genres is read by ebook readers. However, we do not plan on providing an ebook reading service to our customers.

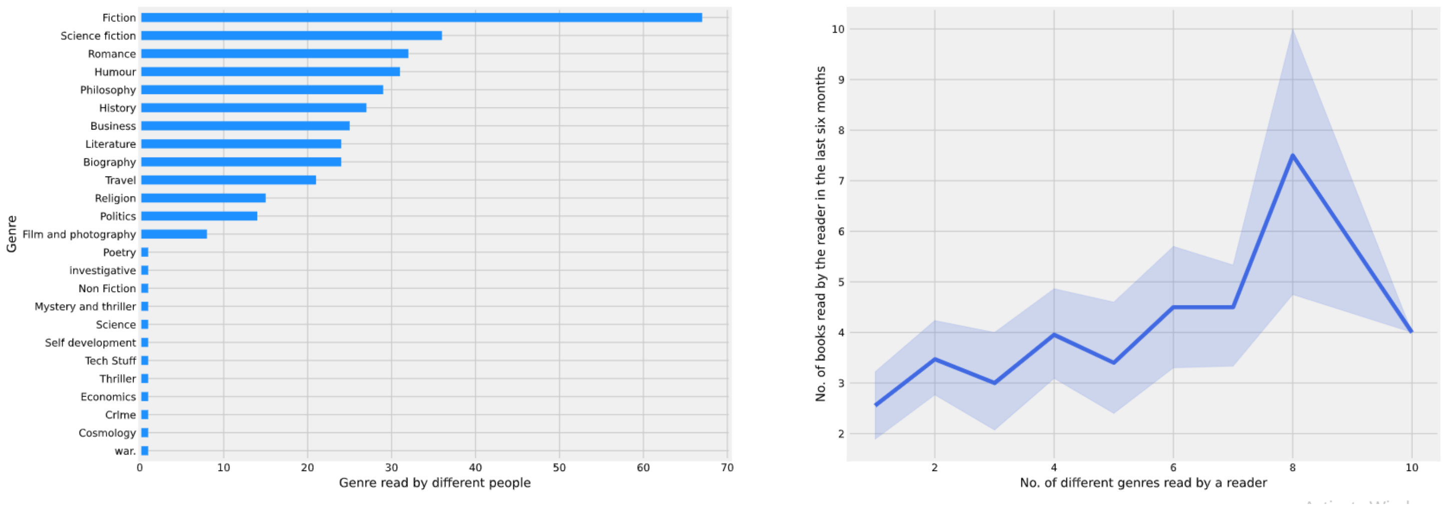
3. The range of books read by paperback and hardcover book readers are the same. We need to probe further to find out the different genres read by them.

4. As expected, people read more no. of paperbacks than hardcover books.

From here onwards all our analysis is on physical book readers.

```
In [3]:  dataf, genres, pphc, d, bv = postprocess(dataf)          #Filters only physical book readers
```

```
In [4]:  plot2(dataf, genres)                                     #Variety
```


Diversity among readers

Observations:

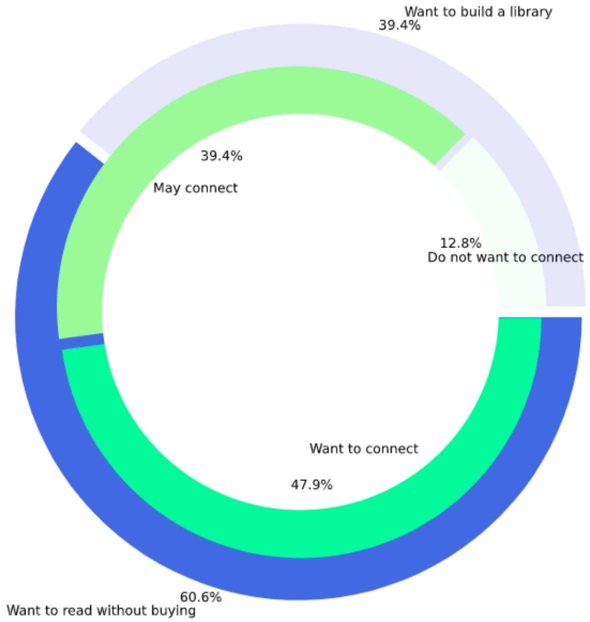
1. The no. of readers for fiction is almost double than that of any other category

2. There are 12 genres which are read by more than 10 people in our survey

3. People who read more no. of books, read a wider range of books.

```
In [5]:  plot3(pphc, d)                                           #Socializing quotient
```

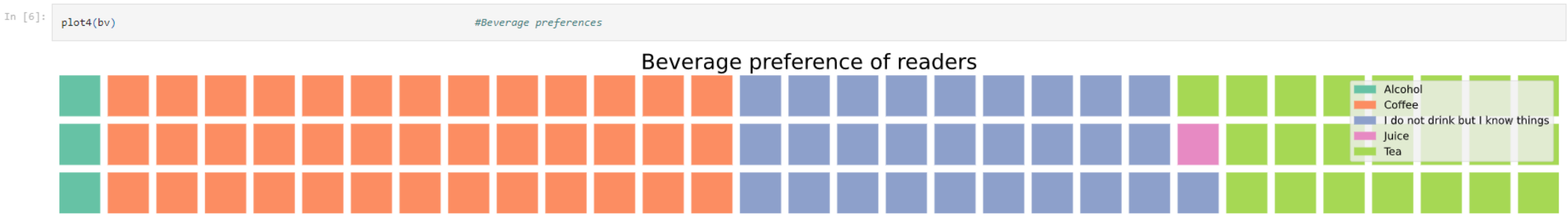
Interest among readers about a book cafe



Observations:

1. A high percentage (60%) of readers want to read books without buying them

Above 85% of readers are open to socializing with other book readers. About 48% readers want to meet fellow book readers



```
In [6]: plot4(bv)                                  #Beverage preferences
```

Beverage preference of readers
Alcohol
Coffee
I do not drink but I know things
Juice
Tea

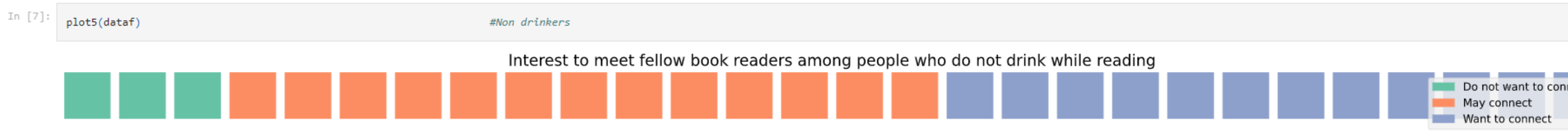
Observations:

1. Maximum readers prefer coffee

2. Most other readers prefer tea or no beverage

Hence providing a reading experience in a cafe is a good idea.

We need to understand how many of the people who do not drink coffee/tea, would want to meet fellow book readers



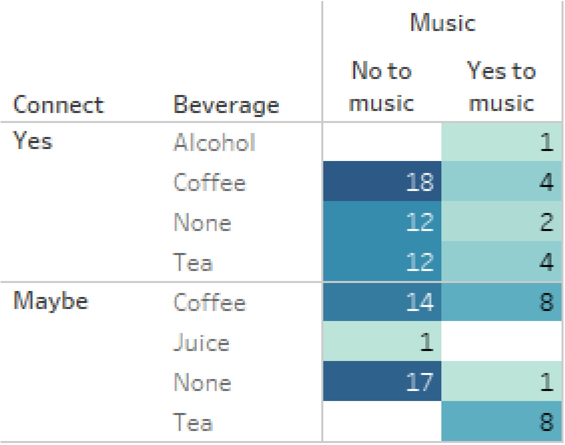
```
In [7]: plot5(dataf)                               #Non drinkers
```

Interest to meet fellow book readers among people who do not drink while reading
Do not want to con
May connect
Want to connect

Observations:

People who do not drink are interested in meeting others. Hence, they would come to meet fellow book readers at the book cafe.

The kind of people we are expecting at out book café:



		Music	
Connect	Beverage	No to music	Yes to music
Yes	Alcohol		1
	Coffee	18	4
	None	12	2
	Tea	12	4
Maybe	Coffee	14	8
	Juice	1	
	None	17	1
	Tea		8

Observations:

1. Most people who want to meet drink coffee
2. Most people do not want to listen to music

Conclusion

1. More than 60% of people are interested in reading books without buying.

2. Above 85% of readers are open to socializing with other book readers.

3. Among expected customers, most are expected to be coffee drinkers.

Future research

1. Most read titles in top ten genres
2. Among people who want to read without buying, why do some people do not want to come to the book cafe?
3. Average reading duration, preferred reading hours and membership options

Code

```
import pandas as pd
import numpy as np
import scipy as s
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
plt.style.use('fivethirtyeight')
pd.set_option('display.max_columns', 500)
from pylab import cm
import warnings
warnings.filterwarnings('ignore')
from pywaffle import Waffle
from copy import deepcopy
from sklearn import feature_selection


def prepare():

    dataf = pd.read_csv(r"C:/Users/Anupam/Downloads/Books_with_coffee.csv")
    dataf.columns = ["Timestamp", "Format", "Beverage", "Music","Target", "Frequency"
,"Connect","Variety"]
    dataf["VarCount"] = dataf.Variety.apply(lambda stri: len(stri.split(",")))   #Cou
nt of variety of books read by people
    format_count = dataf.groupby('Format')['Format'].count()
    return dataf, format_count

def plot1(dataf, format_count):

    fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(22,8))
    sns.countplot(dataf.Format, palette="summer", ax=ax[0])
    ax[0].set_xlabel("Book format")
    ax[0].set_ylabel("Preferred choice of readers")
    sns.boxplot(dataf.Format, dataf.VarCount, palette='summer', ax=ax[1])
    ax[1].set_xlabel("Book format")
    ax[1].set_ylabel("No. of different genres read")
    plt.suptitle("Readership analysis", fontsize=20)
    ax[0].text(-0.05,format_count["Paperback"]+1, s=format_count["Paperback"])
    ax[0].text(0.95,format_count["Hard cover"]+1, s=format_count["Hard cover"])
    ax[0].text(1.95,format_count["Ebook"]+1, s=format_count["Ebook"])
    ax[0].text(2.95,format_count["Audiobook"]+1, s=format_count["Audiobook"])

def postprocess(dataf):

    dataf = dataf[(dataf.Format=="Paperback") | (dataf.Format=="Hard cover")]
    d = dataf[["Variety"]]
           #Only the variety column
    d.Variety = d.Variety.apply(lambda s:s.strip(' ').split(","))
           #Picking up the varieties of genres
    genres = []
    for i in d.Variety:
        for j in i:
            genres.append(j.strip(' '))
           #Storing all the genre varieties in genres
    genres = pd.DataFrame(genres).groupby(0)[0].count()
    genres.drop(['In fiction- fantasy fiction to be exact. Adventure books. Enid Blyt
on( a little kiddish ik). I also would like to read all the Vedas one by one.','None
other than those required for my coursework','Something else'],axis=0,inplace=True) #
Dropping single ultra specific entrie(s)
    genres=genres.sort_values(0)
           #Sorting the genres dataframe

    dataf.Connect.replace({'Yes! no. Well maybe....':'May connect','Yes':'Want to con
nect','No':'Do not want to connect'},inplace=True)
    d = dataf.groupby('Connect')['Connect'].count()
           #Grouping preferences of "Want to connect" attribute
    pphc = pd.DataFrame(dataf.Target)
           #Main column of interest
    pphc.Target.replace({"I wish I could read them without buying a lot of books":"Wa
nt to read without buying",
    "I want to build a library duh!":"Want to build a library"}, inplace=True)
    pphc = pphc.groupby('Target')['Target'].count()
           #Grouping by interest to read books without buying
    dataf.Beverage.replace({'No drink necessary':'I do not drink but I know things','
None':'I do not drink but I know things','No drink necessary':'I do not drink but I k
now things','I drink but not with books':'I do not drink but I know things'},inplace=
True)
    bv =pd.DataFrame(dataf.groupby('Beverage')['Beverage'].count())
           #Grouping by interest of beverage consumption
    bv.drop(['There is no connection between books and beverage'],axis=0,inplace=True
)
    bv.columns=['Count']

    return dataf, genres, pphc, d, bv




def plot2(dataf, genres):
    fig,ax=plt.subplots(nrows=1, ncols=2,figsize=(20,8))
    genres.plot(kind='barh', color="dodgerblue", ax=ax[0])
    #sns.catplot(genres, ax=ax[0])
    plt.suptitle('Diversity among readers', fontsize=20)
    ax[0].set_ylabel('Genre')
    ax[0].set_xlabel('Genre read by different people')
    sns.lineplot(x="VarCount", y="Frequency",data=dataf, ax=ax[1],color='royalblue')
    ax[1].set_xlabel('No. of different genres read by a reader')
    ax[1].set_ylabel('No. of books read by the reader in the last six months')


def plot3(pphc,d):
    plt.figure(figsize=(20,8))
    plt.pie(pphc, autopct='%2.1f%%',colors=['lavender','royalblue'], explode=[0.02,0.02], pctdistance=1.05,labels=pphc.index, labeldistance=1.1)
    plt.title("Interest among readers about a book cafe", fontsize=22)
    #draw circle
    centre_circle = plt.Circle((0,0),0.70,fc='white')
    fig = plt.gcf()
    fig.gca().add_artist(centre_circle)
    # Equal aspect ratio ensures that pie is drawn as a circle
    plt.axis('equal')
    plt.pie(d, radius=0.85, autopct='%1.1f%%', explode=[0.02,0.02,0.02], pctdistance=0.7, labels=d.index, labeldistance=0.55, colors=['mintcream','palegreen','mediumspringgreen'])
    plt.tight_layout()
    plt.show()



def plot4(bv):
    # To plot the waffle Chart
    fig = plt.figure(FigureClass = Waffle, rows = 3, values = bv. Count, labels = list(bv.index) , figsize=(20,8))
    plt.title('Beverage preference of readers',fontsize=20)

def plot5(dataf):
    da = dataf[dataf.Beverage=='I do not drink but I know things']
    x = pd.DataFrame(da.groupby('Connect')['Connect'].count())
    x.columns=['Count']
    fig = plt.figure(FigureClass = Waffle, rows = 1, values = x.Count, labels = list(x.index) , figsize=(20,4))
    plt.title('Interest to meet fellow book readers among people who do not drink while reading')


def plot6(dmod):
    dmod.Beverage.replace({'I do not drink but I know things':'No drink required','None':'No drink required','No drink necessary':'No drink required','I drink but not with books':'No drink required','Depends upon mood and time of day':'No drink required','There is no connection between books and beverage':'No drink required'},inplace=True)
    dmod.Connect.replace({'Yes! no. Well maybe....':'Maybe'},inplace=True)
    dmod.Music.replace({'Yes, like my life':'Yes to music','Nope':'No to music'},inplace=True)
    dk = dmod[["Beverage","Connect","Music"]].groupby(['Connect','Music',"Beverage"])[["Beverage"]].count()
    dk.columns=["Count"]
    dk.sort_values("Connect", ascending=False, inplace=True)
    return dk

def modprocess(datax):

    datax = datax[datax.Beverage!='There is no connection between books and beverage']
    datax.Beverage.replace({'I do not drink but I know things':'No drink required','None':'No drink required','No drink necessary':'No drink required','I drink but not with books':'No drink required','Depends upon mood and time of day':'No drink required'},inplace=True)
    datax.Connect.replace({'Yes':'Yes to connect','No':'No to connect','Yes! no. Well maybe....':'Open to connect'}, inplace=True)
    datax.Music.replace({'Nope':'No to music','Yes, like my life':'Yes to music'},inplace=True)
    datax[list(pd.DataFrame(datax["Format"].unique())[0].sort_values())] = pd.get_dummies(datax.Format)
    datax[list(pd.DataFrame(datax["Beverage"].unique())[0].sort_values())] = pd.get_dummies(datax.Beverage)
    datax[list(pd.DataFrame(datax["Music"].unique())[0].sort_values())] = pd.get_dummies(datax.Music)
    datax.Target.replace({"I wish I could read them without buying a lot of books":1,"I want to build a library duh!":0}, inplace=True)
    datax[list(pd.DataFrame(datax["Connect"].unique())[0].sort_values())] = pd.get_dummies(datax.Connect)
    datax.drop(['Format','Beverage','Music','Connect','Timestamp','Variety'], axis=1, inplace=True)
    datax.drop(['Alcohol','No to connect','Audiobook','Ebook'],axis=1, inplace=True)
    return datax

def plot7(dframe):
    x = dframe.drop(['Target'],axis=1)
    y = dframe.Target
    from sklearn.feature_selection import SelectKBest, chi2, f_classif
    fs = SelectKBest(f_classif, k="all")
    fs.fit(x,y)
    sc = pd.concat([pd.DataFrame(x.columns),pd.DataFrame(fs.scores_)], axis=1)
    sc.columns = ['Feature','Score']
    sc.sort_values('Score', inplace=True, ascending=False)
    plt.figure(figsize=(20,10))
    sns.barplot(sc.Score, sc.Feature, color='seagreen')
    plt.title('Feature importance among people wanting to read books without buying')
```