Homework 2 - CS348 Spring 2019

Description - This assignment is intended to teach you about exploratory data analysis using python and common data science visualization libraries.

Getting Started - You should complete the assignment using your own installation of Python 3 and the packages numpy, pandas, matplotlib, and seaborn. Download the assignment from Moodle and unzip the file. This will create a directory with this file, 'HW02.ipynb', and a 'data' directory. The data files for each data set are in the 'data' directory.

Note: You may need to install the seaborn visualization library. To do this run `conda install seaborn` or `pip install seaborn` in your terminal.

Deliverables - The assignment has a single deliverable: this jupyter notebook file saved as a pdf. Please answer all coding and writing questions in the body of this file. Once all of the answers are complete, download the file by navigating the following menus: File -> Download as -> PDF via LaTeX. Submit the downloaded pdf file on gradescope.

Note: You will be writing the written repsonses in the same cell block as the coding solution, so make sure to comment out the written responses.

Data Sets - In this assignment, you will conduct an exploratory data analysis on 2 datasets. The first dataset, 'flights', is imported for you from the seaborn library.

Academic Honesty Statement — Copying solutions from external sources (books, web pages, etc.) or other students is considered cheating. Sharing your solutions with other students is considered cheating. Posting your code to public repositories such as GitHub is also considered cheating. Any detected cheating will result in a grade of 0 on the assignment for all students involved, and potentially a grade of F in the course.

This academic honesty statement does not restrict you from reading official documentation or using other web resources for understanding the syntax of python, related data science libraries, or properties of distributions.

In [68]:

```
# Do not import any other libraries other than those listed here.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Problem 1 - Flights Dataset

In this problem you'll analyze a dataset of the monthly number of passengers on US flights from the years 1949 - 1960.

Part 1 (3 points)
Load the 'flights' dataset as a pandas DataFrame and print the first 10 rows of data.

In [69]:

```
# Part 1 Solution

# --- write code here ---
d=pd.read_csv("data/flights.csv", sep=',', na_values=[' ?'], engine='python').iloc[:, 1:]
print(d.head(10))
```

```
       month  passengers
0    January         112
1   February         118
2      March         132
3      April         129
4        May         121
5       June         135
6       July         148
7     August         148
8  September         136
9    October         119
```

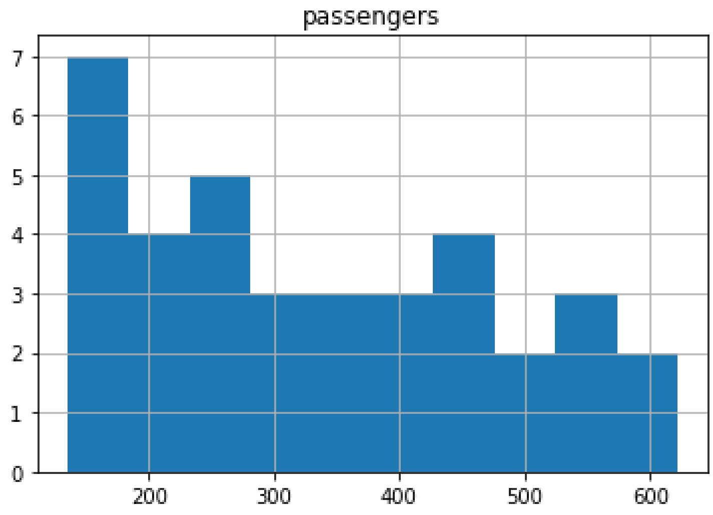
Part 2 (12 points)
Plot a histogram of the number of montly passengers traveling during the summer months. Summer months include June, July, and August. Does the histogram resemble a normal distribution? Provide two justifications for your answer.

In [70]:

```
# Part 2 Solution

# --- write code here ---
sum_month=d.loc[(d['month']=='June') | (d['month']=='July') | (d['month']=='August')]
sum_month.hist(column='passengers')
print(sum_month.head(36))
# --- written response here ---
# The histogram doesn't resemble a normal distribution.
# 1. The number of passengers of summer months increases each year, which is not normal distribution.
# 2. The mean, mode and median are different, thus it's not normal distribution.
```

```
       month  passengers
5       June         135
6       July         148
7     August         148
17      June         149
18      July         170
19    August         170
29      June         178
30      July         199
31    August         199
41      June         218
42      July         230
43    August         242
53      June         243
54      July         264
55    August         272
65      June         264
66      July         302
67    August         293
77      June         315
78      July         364
79    August         347
89      June         374
90      July         413
91    August         405
101     June         422
102     July         465
103   August         467
113     June         435
114     July         491
115   August         505
125     June         472
126     July         548
127   August         559
137     June         535
138     July         622
139   August         606
```

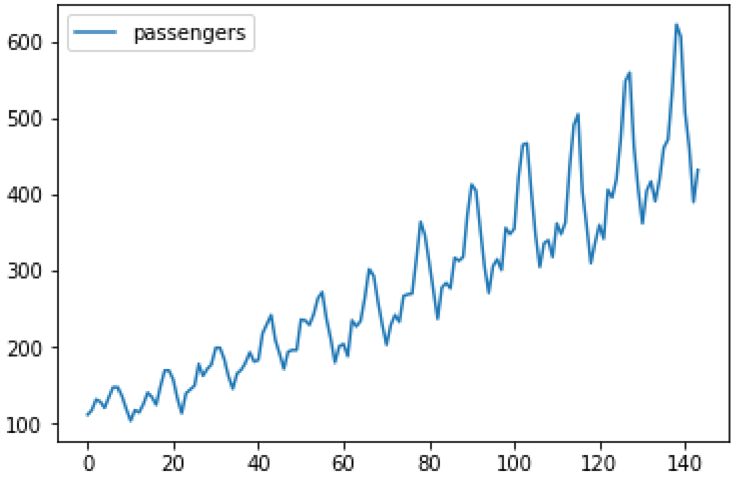


Part 3 (12 points)
Make a timeseries plot using `sns.lineplot()` . Using this plot, answer the following question: are the rows in the dataset independent and identically distributed (IID)? Why or why not?

In [71]:

```
# Part 3 Solution

# --- write code here ---
new_d = d.drop(["month"], axis = 1)
plot = sns.lineplot(data=new_d)
# --- written response here ---
# --- I dropped the 'month' column since it's not numeric values. ---
# --- 1. As the graph shows, the number of passengers increases in the f
irst half year and then decreases
# --- in the second half year, which means it's affected by time. Thus,
 it's not idependent. ---
# --- 2. The number of passengers of each month has a trend of increasin
g. Thus, it's not identically distributed. ---
```



Part 4 (20 points) Your colleague is trying to predict the total number of passengers who flew in June 1961, one year after the end of the flights dataset. They notice that the average difference between the number of passengers in June and the number passengers in January is 69.92 from 1949 - 1960. Given that they know that 450 passengers flew in January 1961, they predict that there will be 520 passengers flying in June 1961.

Is this a good estimate for the number of passengers flying in June 1961? If not, do you expect it to over- or under-estimate the actual number of passengers? Explain your answer.

In [72]:
```
# Part 3 Solution

# --- written response here ---
# --- 1. Using mean of difference is not a good approach to estimate the
 number of passengers flying in June 1961
# --- because it could be affected by outliers.
# --- 2. By analyzing the dataset we could know that the number of passe
ngers flying in June increases each year. ---
# --- 3. By the end of 1960, the number is 535, and with the trend shows
 us the number will be higher than 535 in 1961
# --- we then know the estimation is not accurate, and it's under-estima
te the actual number of passengers. ---
```

Problem 2 - Synthetic Data

In this problem you'll be asked to analyze a synthetic dataset of four variables.

Part 1 (3 points)
Load the 'synthetic' data as a pandas dataframe and print the first 10 rows.

In [73]:
```
# Part 1 Solution

# --- write code here ---
dataset=pd.read_csv("data/synthetic.csv", sep=',', na_values=[' ?'], eng
ine='python').iloc[:, 1:]
print(dataset.head(10))
```

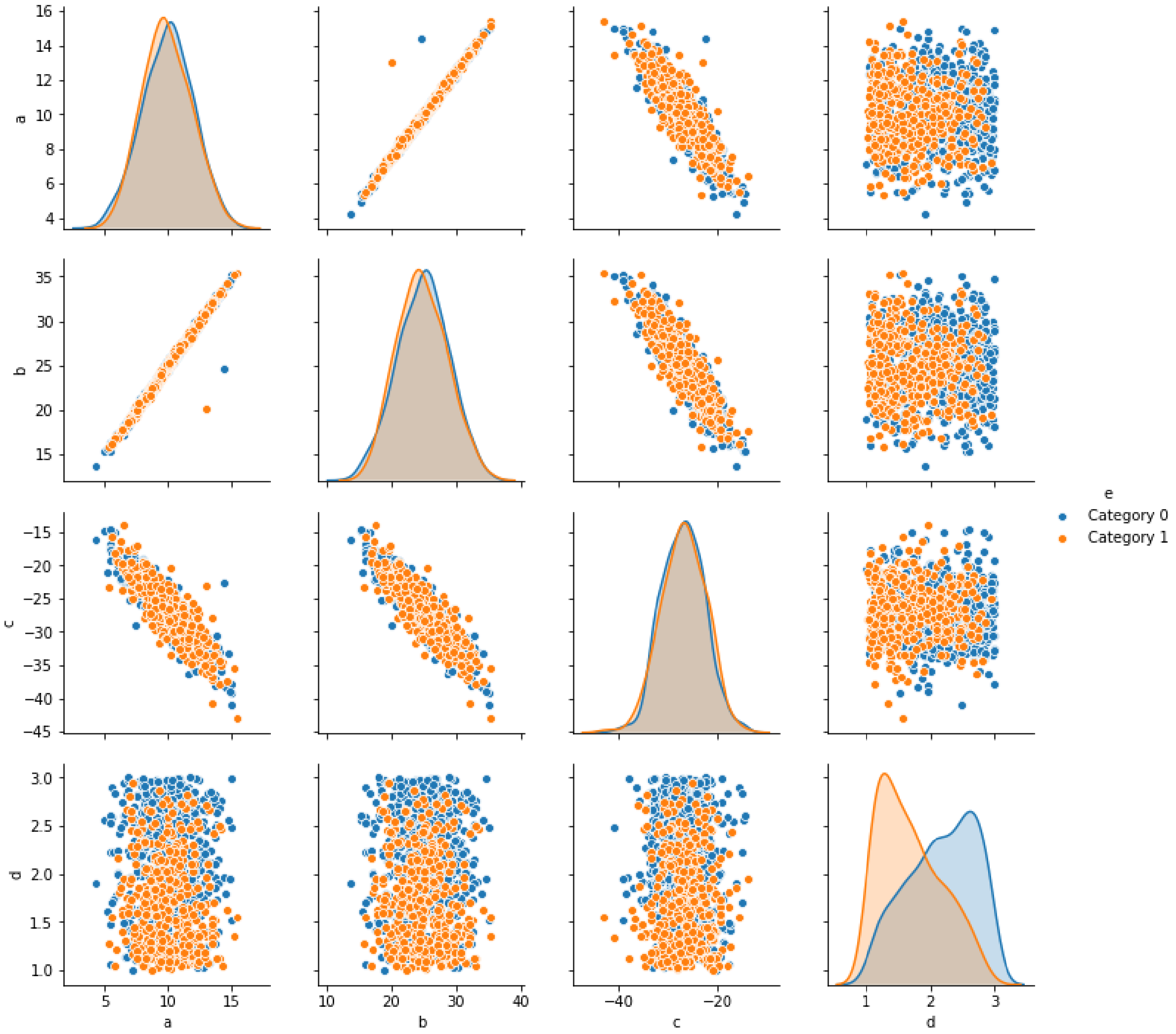
```
           a          b          c         d           e
0  10.590051  26.191885 -30.634067  2.364324  Category 0
1   8.669480  22.364265 -25.528595  1.558527  Category 0
2  10.473553  26.396606 -25.777483  1.806674  Category 0
3   5.946901  17.069787 -19.311699  1.207391  Category 1
4  11.449457  28.229418 -28.239489  1.054686  Category 1
5   9.160058  23.213218 -24.035343  1.232872  Category 1
6  11.768469  28.725059 -31.058865  2.171159  Category 0
7   7.982466  20.984785 -23.001933  2.712406  Category 0
8   9.711183  24.984666 -25.254299  1.124860  Category 0
9   9.705916  24.025878 -22.038090  2.683041  Category 0
```

Part 2 (6 points)
Use `sns.pairplot()` to make a pairplot of the synthetic dataset. Use colors and/or symbols to visualize the relationship between the categorical and ordinal variables.

```
In [74]: # Part 2 Solution

         # --- write code here ---
         sns.pairplot(dataset, hue="e")
```

```
Out[74]: <seaborn.axisgrid.PairGrid at 0x1a1979f7b8>
```



Part 3 (10 points)
For variables a, b, c, and d determine whether its marginal distribution is uniform or normal. Explain your answers.

Hint: Be careful about the `diag_kind` parameter in `sns.lineplot`. The default behaviour is to use a smoothed estimate of the probability density that can sometimes be misleading. None of the marginal distributions are a mixture of normal distributions.

In [75]:

```
# Part 3 solution

# --- written response here ---
# --- 1. a, b, c have normal marginal distributions base on the graphs a
re in bell shape. ---
# --- 2. d has a uniform marginal distribution since the graph is not di
stinguished by e. ---
# --- And the graph shape are about at the same height. ---
```

Part 4 (10 points)
For all pairs of variables (a-b, a-c, etc.) in the synthetic dataset, determine if the two variables are marginally independent. Explain your answers.

Reminder: A random variable X is marginally independent of another random variable Y if knowledge of X does not change the distribution of Y. In other words, X and Y are marginally independent if P(X,Y) = P(X)P(Y) or equivalently P(Y|X) = P(Y).

In [76]:

```
# Part 4 solution

# --- written response here ---
# --- a-b, a-c is not marginally independent, the plot is close to linea
r regression. ---
# --- a-d is marginally independent, since the plots are randomly distri
buted. ---
# --- a-e is marginally independent, since change of a doesn't affect va
lue of e. ---
# --- b-c is not marginally independent, the plot is close to linear reg
ression. ---
# --- b-d is marginally independent, since the plots are randomly distri
buted. ---
# --- b-e is marginally independent, since change of b doesn't affect va
lue of e. ---
# --- c-d is marginally independent, since the plots are randomly distri
buted. ---
# --- c-e is marginally independent, since change of c doesn't affect va
lue of e. ---
# --- d-e is marginally independent, since change of d doesn't affect va
lue of e. ---
```

Part 5 (12 points)
There are 2 outliers in the synthetic dataset. Is it possible to identify these outliers using only the marginal distributions for each of the 4 variables? Why or why not?

In [77]:

```
# Part 5 solution

# --- written response here ---
# --- It's impossible to identify outliers using only the marginal distr
ibutions. ---
# --- Since the distribution shows the number in each intervals instead
 of unqie values. ---
```

Part 6 (12 points)

Given your answer to part 5, is it plausible to detect outliers using only visualizations when your dataset contains 1000 columns? Why or why not?

In [78]:

```
# Part 6 solution

# --- written response here ---
# --- For large amount of data, visualization could be ineffective. ---
# --- Given answer to part 5, we could easily detect the outliers in a-b
 or b-a graphs. ---
# --- However, in large amount of columns, we need to generate more grap
hs to show relations between variables. ---
# --- To detect outliers in such big amount of graphs is inefficient. --
-
```