CMPINF 2100: Midterm exam

YOUR NAME HERE !!!!

Assigned: Tuesday of Week 09 at 11:00PM

DUE: Tuesday of Week 10 at 11:59PM

You may add as many code and markdown cells as you see fit to answer the questions.

You are NOT allowed to collaborate with anyone on this exam.

Overview

You will demonstrate your ability to merge, group, summarize, visualize, and find patterns in data. This exam uses data associated with a manufacturing example. An overview of the goals, considerations, CSV files, and variables within the data is provided in a presentation on Canvas. Please read through those slides before starting the exam.

The data are provided in 5 separate CSV files. The CSV files are available on Canvas. You **MUST** download the files and save them to the same working directory as this notebook.

The specific instructions in this notebook tell you when you must JOIN the data together. Please read the problems carefully.

The overall objective of this exam is to JOIN data from multiple files in order to explore and find interesting patterns between the machine operating conditions and supplier information. You will report your findings within this notebook by displaying Pandas DataFrames and statistical visualizations via Seaborn and matplotlib when necessary.

Import modules

You are permitted to use the following modules on this exam.

In [ ]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

import seaborn as sns
```

You may also use the following functions from scikit-learn on this exam.

```
In [ ]: from sklearn.preprocessing import StandardScaler
        from sklearn.cluster import KMeans
        from sklearn.decomposition import PCA
```

You may also use the following sub module from SCIPY.

```
In [ ]: from scipy.cluster import hierarchy
```

You are **NOT** permitted to use any other modules or functions. However, you **ARE** permitted to create your own user defined functions if you would like.

Problem 01

The file names for the 3 machine data sets are provided as strings in the cell below. You are required to read in the CSV files and assign the data to the `m01_df`, `m02_df`, and `m03_df` objects. The data from machine 1 will therefore be associated with `m01_df`, machine 2 is associated with `m02_df`, and machine 3 is associated with `m03_df`.

In this problem you must explore each of the three machine data sets.

You must perform the following **ESSENTIAL** activities:

- How many rows and columns are in each data set?
- What are the names and data types for each column?
- How many unique values are there for each column?
- How many missing values are there for each column?

You must visually explore the MARGINAL behavior of the variables in the data. You must use visualizations appropriate for the DATA TYPE of the columns.

You must visually explore RELATIONSHIPS between variables in the data. You must use visualizations appropriate for the DATA TYPES. You must make sure that your visualizations can answer the following questions:

- How many unique values for `Batch` are associated with each MACHINE (data set)?
- How many cell phone cases are associated with each `Batch` value for each MACHINE (data set)?
- Do the summary statistics of the OPERATING VARIABLES `x1` through `x4` vary across the three MACHINES?
- Do the summary statistics of the OPERATING VARIABLES `x1` through `x4` vary across the `Batch` values?
- Do the relationships between the OPERATING VARIABLES `x1` through `x4` vary across the three MACHINES?
- Do the relationships between the OPERATING VARIABLES `x1` through `x4` vary across the `Batch` values?

At the conclusion of this problem, you **MUST** CONCATENATE the 3 MACHINE data sets into a single DataFrame. The single DataFrame must be named `machine_df`. Before concatenating, you **MUST** add a column `machine_id` to each DataFrame with the correct index value for that machine (1, 2, or 3). The concatenating DataFrame variable name is provided as a reminder to you below.

You may add as many markdown and code cells as you see fit to answer this question. Include markdown cells stating what you see in the figures and why you selected to use them.

SOLUTION

In [ ]:

```
# Define the files's for the 3 machine level CSV files

file_m01 = 'midterm_machine_01.csv'

file_m02 = 'midterm_machine_02.csv'

file_m03 = 'midterm_machine_03.csv'
```

In [ ]:

```
# read in the CSV files and name them accordingly

m01_df =

m02_df =

m03_df =
```

In [ ]:

```
# concatenate the 3 DataFrames into a single DataFrame which includes the `mac

machine_df =
```

Problem 02

The supplier batch data set file name is provided for you below. You must read in the CSV file and assign the data set to the `batch_df` object.

You must perform the following **ESSENTIAL** activities:

- How many rows and columns are in the data?
- What are the names and data types for each column?
- How many unique values are there for each column?
- How many missing values are there for each column?

You must visually explore the MARGINAL behavior of the variables in the data. You must use visualizations appropriate for the DATA TYPE of the columns.

You must visually explore RELATIONSHIPS between variables in the data. You must use visualizations appropriate for the DATA TYPES. You must make sure that your visualizations can answer the following questions:

- Do the summary statistics for `Density` depend on the `Supplier` ?
- Does the average `Density` depend on the `Supplier` ?
- How does `Density` relate to `Batch` for each `Supplier` ?

After exploring the `batch_df` DataFrame, you **MUST** JOIN/MERGE the `batch_df` DataFrame with the `machine_df` DataFrame. Assign the merged DataFrame to the `dfa` DataFrame.

You can now explore the relationships between the MACHINE OPERATIONAL VARIABLES and the SUPPLIERS! You must use visualizations to explore the following relationships:

- Explore if the summary statistics of the 4 OPERATING VARIABLES `x1` through `x4` vary across `Batch` for each MACHINE given each `Supplier` . Your figures MUST use `Batch` as the x-axis variable.
- Explore if the relationships between the 4 OPERATING VARIABLES `x1` through `x4` vary across `Supplier` .

You may add as many markdown and code cells as you see fit to answer this question.

SOLUTION

In [ ]:

```
# define the batch supplier file
batch_file = 'midterm_supplier.csv'
```

In [ ]:

```
# read in the batch supplier data set

batch_df =
```

In [ ]:

```
# merge the batch supplier data set with the (concatenated) machine data set

dfa =
```

Problem 03

The DROP TEST result data set file name is provided for you below. You must read in the CSV file and assign the dta set to the `test_df` object.

You must perform the following **ESSENTIAL** activities:

- How many rows and columns are in the data?
- What are the names and data types for each column?
- How many unique values are there for each column?
- How many missing values are there for each column?

You must visually explore the MARGINAL behavior of the variables in the data. You must use visualizations appropriate for the DATA TYPE of the columns.

You must visually explore RELATIONSHIPS between variables in the data. You must use visualizations appropriate for the DATA TYPES. You must make sure that your visualizations can answer the following questions:

- Count the number of times each unique value of `Result` occurs for each `test_group_id` value.

After exploring the `test_df` DataFrame, you **MUST** JOIN/MERGE the `test_df` DataFrame with the `dfa` DataFrame. Assign the merged DataFrame to the `dfb` DataFrame. You **MUST** answer the following:

- How many rows remain using the DEFAULT joining procedure?

You may add as many markdown and code cells as you see fit to answer this question.

SOLUTION

In [ ]:

```
# define the test data set file name
test_file = 'midterm_test.csv'
```

In [ ]:

```
# read in the test data set

test_df = 
```

In [ ]:

```
# merge test_df with the dfa object

dfb = 
```

Problem 04

You must now examine the merged `dfb` object and answer the following:

- Count the number of times each unique value of `Result` occurs for each value of `machine_id`.
- Count the number of times each unique value of `Result` occurs for each value of `Supplier`.
- Visualize the number of times each unique value of `Result` occurs per `Batch` for each value of `machine_id`.
- Visualize the number of times each unique value of `Result` occurs per `Batch` for each value of `machine_id` and `Supplier`.
- Calculate the PROPORTION of times the cell phone case failed the test in each `Batch` per `machine_id`.
- Visualize the PROPORTION of times the cell phone case failed the test in each `Batch` per `machine_id` and for each unique value of `Supplier`.

HINT: Remember that a FAILED test is encoded as `Result == 1`. How can you calculate the PROPORTION of times `Result == 1`?

Add as many cells as you see fit to answer this question.

SOLUTION

```
In [ ]:
```

Problem 05

You must cluster the rows of `dfb` using the 4 operational variables `x1` through `x4`. You must decide how many clusters to use and describe how you made that choice. You may use KMeans OR Hierarchical clustering. Include any figures that helped you make that choice.

Visualize your cluster analysis results by:

- Plotting the number of observations per cluster.
- Visualizing the relationships between the operational variables GIVEN the cluster.

You are interested in the PROPORTION of cell phone cases that failed the DROP TEST. Are any of the clusters associated with higher failure PROPORTIONS than others? Based on your visualizations how would you describe that cluster?

Add as many cells as you see fit to answer this question.

SOLUTION

```
In [ ]:
```