Lab_5_Andrade

February 4, 2024

14 Febuary 2024 #

Lab 5 Assignment - CS 4315

Doug Andrade

1. Load the armada.csv into a Pandas dataframe.

```
[46]: # Import the Pandas module for data frame operations
      import pandas as pd

      # Set the csv file name to a object, to avoid hard-coding
      csv_file = 'armada.csv'

      # Read-in the csv file as a Pandas data frame, as an object to be operated on␣
        ↪later
      armada_df = pd.read_csv(filepath_or_buffer = csv_file)

      armada_df.head()
```

```
[46]:            Battle  Year  Portuguese Ships  Dutch Ships  English Ships  \
      0          Bantam  1601                 6            3              0
      1  Malacca Strait  1606                14           11              0
      2   Ilha das Naus  1606                 6            9              0
      3      Pulo Butum  1606                 7            9              0
      4          Surrat  1615                 6            0              4

         Ratio of Portuguese Ships to Dutch/British Ships Spanish Involvement  \
      0                                             2.000                  No
      1                                             1.273                  No
      2                                             0.667                  No
      3                                             0.778                  No
      4                                             1.500                  No

        Portuguese Outcome
      0               Draw
      1               Draw
      2             Defeat
      3            Victory
```

```
4                Draw
```

2. Get the mean, standard deviation, max, and min of each numeric column.

[2]:

```
# the describe() function quickly outputs statistical metrics for the numeric␣
 ↪features (only) of the data set
armada_df.describe()
```

[2]:

```
              Year  Portuguese Ships  Dutch Ships  English Ships  \
count    28.000000         28.000000    28.000000      28.000000
mean   1628.392857         13.142857    13.428571       1.785714
std      17.559084         15.922763    22.280083       5.927695
min    1588.000000          2.000000     0.000000       0.000000
25%    1618.750000          5.750000     4.000000       0.000000
50%    1628.500000          6.000000     8.000000       0.000000
75%    1639.000000         14.000000    11.000000       0.000000
max    1658.000000         69.000000   110.000000      31.000000

       Ratio of Portuguese Ships to Dutch/British Ships
count                                         28.000000
mean                                           1.159893
std                                            0.928341
min                                            0.150000
25%                                            0.650250
50%                                            0.928500
75%                                            1.500000
max                                            4.636000
```

3. Create z-scored columns for each numeric column using the statistics for the samples given.

[3]:

```
# Using list comprehension and conditional statements, create a list of all the␣
 ↪numeric column names
numeric_col_list = [col for col in armada_df.columns if armada_df[col].dtype !=␣
 ↪'object' and col != 'Year']

# Iterate only the data frame columns that exist in the "numerica_col_list"␣
 ↪above
for col in numeric_col_list:
    # Create a new column name for the cooresponding columns Z-Score
    zscored_col_name = col + ' Z-Score'
    # Immediately insert the new Z-Score column next to the parent column by␣
 ↪adding one to the parent's column position
    # within the insert function
    armada_df.insert(armada_df.columns.get_loc(col) + 1,
                     # Calculate the Z-Score (x - x_bar) / x_std)
                     zscored_col_name, (armada_df[col] - armada_df[col].mean())␣
 ↪/ armada_df[col].std())


armada_df.head()
```

[3]:

```
           Battle  Year  Portuguese Ships  Portuguese Ships Z-Score  \
0          Bantam  1601                 6                 -0.448594   
1  Malacca Strait  1606                14                  0.053831   
2   Ilha das Naus  1606                 6                 -0.448594   
3      Pulo Butum  1606                 7                 -0.385791   
4          Surrat  1615                 6                 -0.448594   

   Dutch Ships  Dutch Ships Z-Score  English Ships  English Ships Z-Score  \
0            3            -0.468067              0              -0.301249   
1           11            -0.109002              0              -0.301249   
2            9            -0.198768              0              -0.301249   
3            9            -0.198768              0              -0.301249   
4            0            -0.602716              4               0.373549   

   Ratio of Portuguese Ships to Dutch/British Ships  \
0                                            2.000   
1                                            1.273   
2                                            0.667   
3                                            0.778   
4                                            1.500   

   Ratio of Portuguese Ships to Dutch/British Ships Z-Score  \
0                                           0.904955          
1                                           0.121838          
2                                          -0.530939          
3                                          -0.411371          
4                                           0.366360          

  Spanish Involvement Portuguese Outcome
0                  No               Draw
1                  No               Draw
2                  No             Defeat
3                  No            Victory
4                  No               Draw
```

4. Create either a binary column or one-hot columns for the non-numeric columns.

[17]:

```
# Identify non-numeric columns
non_numeric_col_list = armada_df.select_dtypes(exclude = 'number').columns
```

```
# Create a new DataFrame with one-hot encoded columns for non-numeric columns,␣
 ↪concated along the column axis
# I kept the parameter "drop_first" as False for completeness only, knowning␣
 ↪there would be unintended correlation
armada_dummies = pd.concat(objs = [armada_df, pd.get_dummies(data =␣
 ↪armada_df[non_numeric_col_list],
                                                             drop_first =␣
 ↪False)],
                a axis = 1)

armada_dummies.head()
```

[17]:

```
           Battle  Year  Portuguese Ships  Portuguese Ships Z-Score  \
0          Bantam  1601                 6                 -0.448594   
1  Malacca Strait  1606                14                  0.053831   
2   Ilha das Naus  1606                 6                 -0.448594   
3      Pulo Butum  1606                 7                 -0.385791   
4          Surrat  1615                 6                 -0.448594   

   Dutch Ships  Dutch Ships Z-Score  English Ships  English Ships Z-Score  \
0            3            -0.468067              0              -0.301249   
1           11            -0.109002              0              -0.301249   
2            9            -0.198768              0              -0.301249   
3            9            -0.198768              0              -0.301249   
4            0            -0.602716              4               0.373549   

   Ratio of Portuguese Ships to Dutch/British Ships  \
0                                             2.000   
1                                             1.273   
2                                             0.667   
3                                             0.778   
4                                             1.500   

   Ratio of Portuguese Ships to Dutch/British Ships Z-Score  …  \
0                                           0.904955         …   
1                                           0.121838         …   
2                                          -0.530939         …   
3                                          -0.411371         …   
4                                           0.366360         …   

  Battle_Pulo Butum Battle_Recife  Battle_Surrat  Battle_Tamandare  \
0             False         False          False             False   
1             False         False          False             False   
2             False         False          False             False   
3              True         False          False             False   
4             False         False           True             False   
```

```
   Spanish Involvement_?  Spanish Involvement_No  Spanish Involvement_Yes  \
0                  False                    True                    False
1                  False                    True                    False
2                  False                    True                    False
3                  False                    True                    False
4                  False                    True                    False

   Portuguese Outcome_Defeat  Portuguese Outcome_Draw  \
0                      False                     True
1                      False                     True
2                       True                    False
3                      False                    False
4                      False                     True

   Portuguese Outcome_Victory
0                       False
1                       False
2                       False
3                        True
4                       False

[5 rows x 36 columns]
```

5. Calculate and display, with a color gradient, the correlation between the z-scored, binary, and one-hot columns, excluding the battle one-hot columns.

[45]:

```
# List the columns to calculate and display correlation as an object
cols_to_corr = ['Portuguese Ships Z-Score',
                'Dutch Ships Z-Score',
                'English Ships Z-Score',
                'Ratio of Portuguese Ships to Dutch/British Ships Z-Score',
                'Spanish Involvement_?',
                'Spanish Involvement_No',
                'Spanish Involvement_Yes',
                'Portuguese Outcome_Defeat',
                'Portuguese Outcome_Draw',
                'Portuguese Outcome_Victory',
               ]

# run the Pandas .corr() function with the bwr matplotlib colormap for greater␣
 ↪contrast
armada_dummies[cols_to_corr].corr().style.background_gradient(cmap = 'bwr',
                                                              vmin = -1.0, #␣
 ↪min colormap value
                                                              vmax = 1.0) # max␣
 ↪colormap value
```

```
[45]: <pandas.io.formats.style.Styler at 0x7f289ea9a190>
```