HO CHI MINH UNIVERSITY OF SCIENCE

MULTIVARIATE STATISTICS ANALYSIS - LAB03

March 25, 2024

Author

Nguyen Viet Kim - 21127333

Contents

1 Introduction

The program use Scikitlearn for calculating the basic statistics for multivariate data. The program use the wine.csv file for input data.

```
## Data:
     V1     V2    V3    V4    V5   V6    V7    V8    V9   V10    V11   V12
0     1  13.20  1.78  2.14  11.2  100  2.65  2.76  0.26  1.28   4.38  1.05
1     1  13.16  2.36  2.67  18.6  101  2.80  3.24  0.30  2.81   5.68  1.03
2     1  14.37  1.95  2.50  16.8  113  3.85  3.49  0.24  2.18   7.80  0.86
3     1  13.24  2.59  2.87  21.0  118  2.80  2.69  0.39  1.82   4.32  1.04
4     1  14.20  1.76  2.45  15.2  112  3.27  3.39  0.34  1.97   6.75  1.05
..   ..    ...   ...   ...   ...  ...   ...   ...   ...   ...    ...   ...
172   3  13.71  5.65  2.45  20.5   95  1.68  0.61  0.52  1.06   7.70  0.64
173   3  13.40  3.91  2.48  23.0  102  1.80  0.75  0.43  1.41   7.30  0.70
174   3  13.27  4.28  2.26  20.0  120  1.59  0.69  0.43  1.35  10.20  0.59
175   3  13.17  2.59  2.37  20.0  120  1.65  0.68  0.53  1.46   9.30  0.60
176   3  14.13  4.10  2.74  24.5   96  2.05  0.76  0.56  1.35   9.20  0.61

      V13   V14
0    3.40  1050
1    3.17  1185
2    3.45  1480
3    2.93   735
4    2.85  1450
..    ...   ...
172  1.74   740
173  1.56   750
174  1.56   835
175  1.62   840
...
 12  V13     177 non-null    float64
 13  V14     177 non-null    int64
dtypes: float64(11), int64(3)
memory usage: 19.5 KB
```

Figure 1: Data information

2 Importing Libraries

- **matplotlib.pyplot**: This library is used for creating visualizations such as scatter plots and histograms.
- **numpy**: It provides support for mathematical operations on arrays and matrices, essential for numerical computing.
- **pandas**: Pandas is a powerful library for data manipulation and analysis, particularly useful for handling structured data.
- **scipy.stats**: This module from SciPy provides various statistical functions, including correlation and hypothesis testing.
- **seaborn**: Seaborn is a statistical data visualization library built on top of matplotlib, offering enhanced aesthetics and additional plot types.
- **sklearn.preprocessing**: This module from scikit-learn provides functions for preprocessing data, such as scaling and normalization.

- **sklearn.decomposition**: It includes methods for decomposition techniques like Principal Component Analysis (PCA).

3 Data Visualization

Scatter Matrix Plot: The scatter matrix plot visualizes the relationships between pairs of variables in the dataset, while kernel density estimation (KDE) plots along the diagonal show the distribution of each variable.

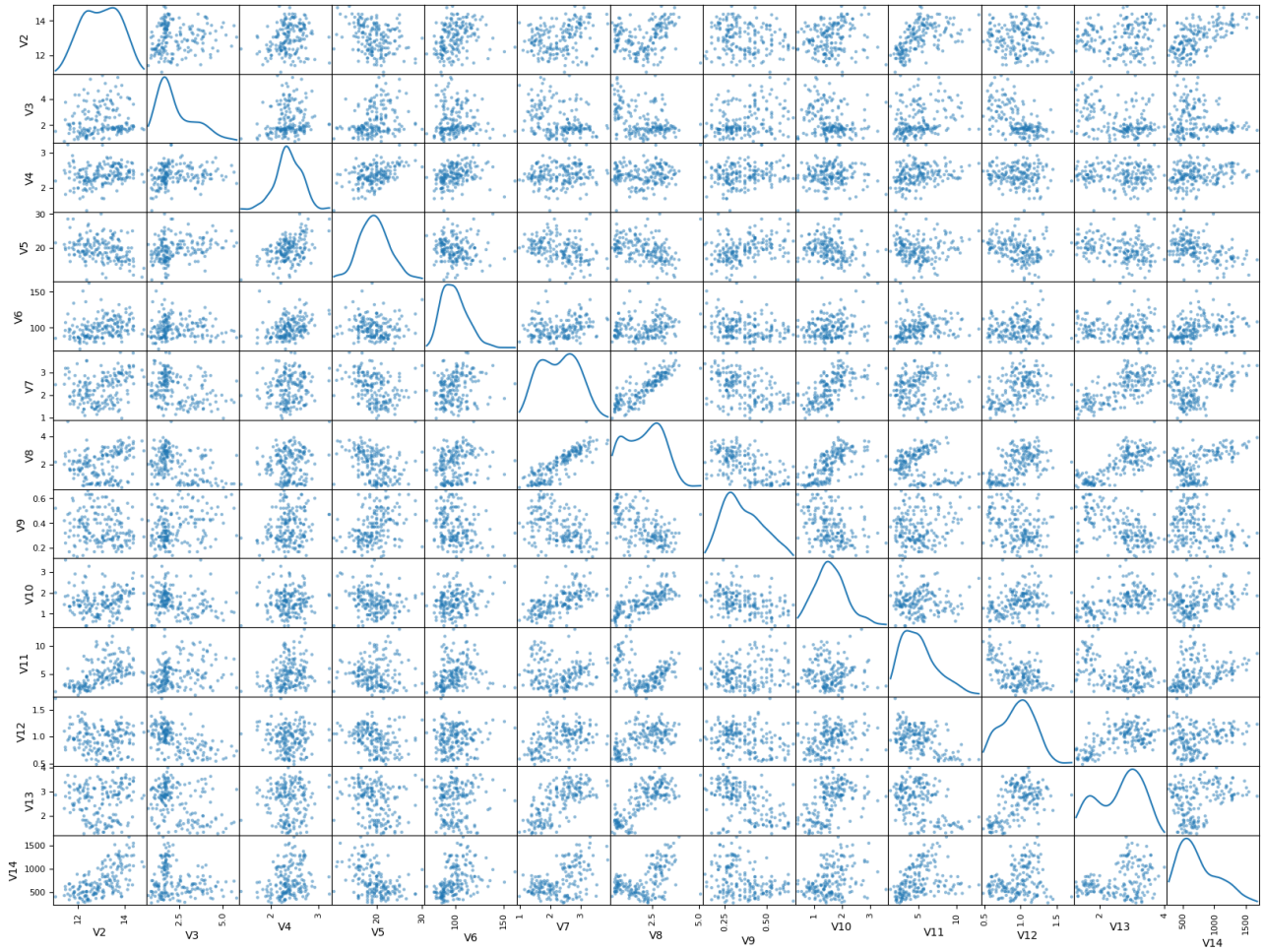

Figure 2: Scatter matrix plot

LM Plots and Line Plots: LM plots (scatter plots with overlaid regression lines) and line plots are used to visualize relationships between specific pairs of variables and trends over time, respectively.

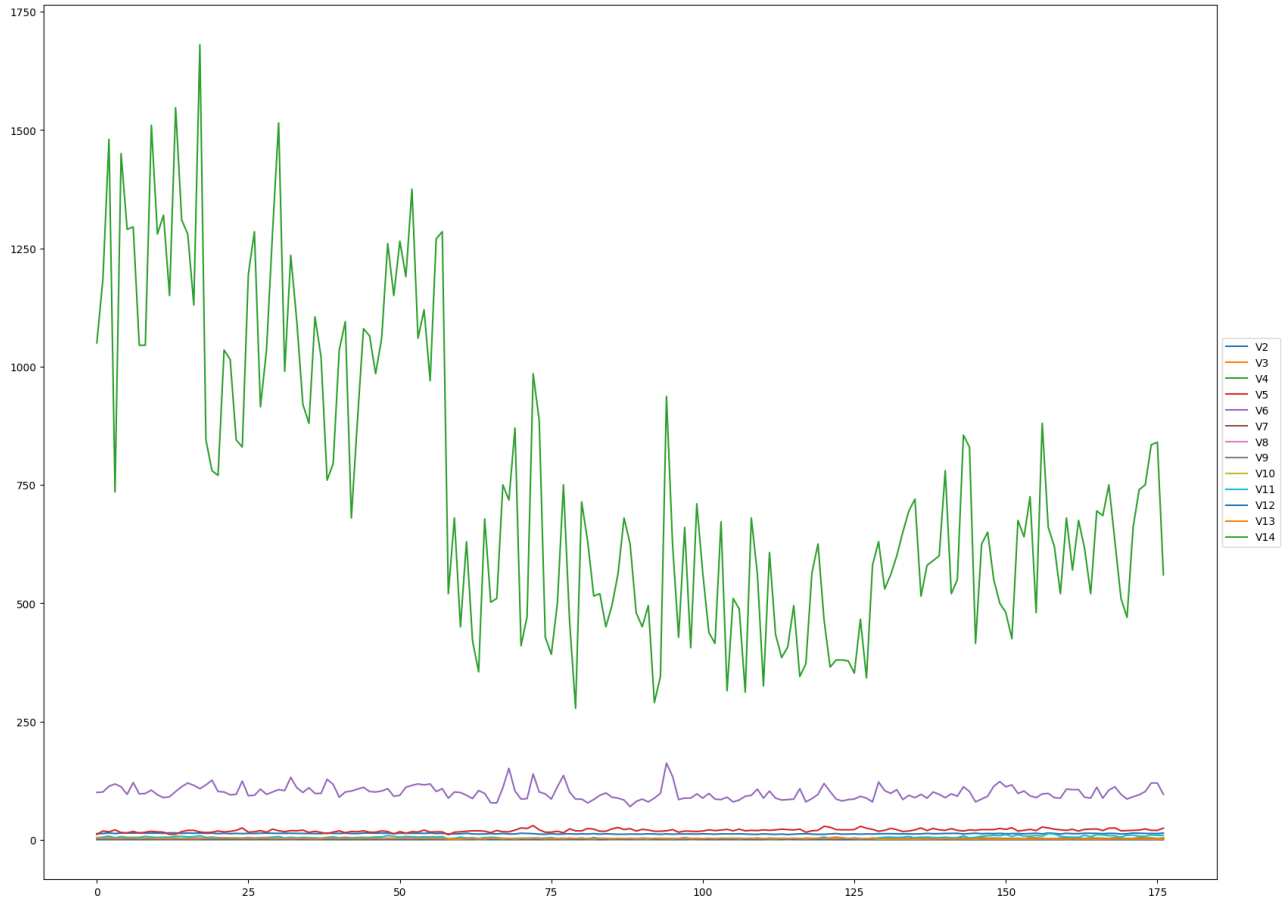

Figure 3: LM plot and Line plot

4 Summary Statistics Calculation

Mean, Standard Deviation, Maximum, Minimum: These statistics are computed for each variable in the dataset using numpy functions (`np.mean()`, `np.std()`, `np.max()`, `np.min()`), providing insights into the central tendency and variability of the data.

	Mean	Standard Deviation	Maximum	Minimum
V2	12.993672	0.806520	14.83	11.03
V3	2.339887	1.116148	5.80	0.74
V4	2.366158	0.274302	3.23	1.36
V5	19.516949	3.326634	30.00	10.60
V6	99.587571	14.133922	162.00	70.00
V7	2.292260	0.624693	3.88	0.98
V8	2.023446	0.995833	5.08	0.34
V9	0.362316	0.124300	0.66	0.13
V10	1.586949	0.569928	3.58	0.41
V11	5.054802	2.317871	13.00	1.28
V12	0.956983	0.228487	1.71	0.48
V13	2.604294	0.703108	4.00	1.27
V14	745.096045	313.993283	1680.00	278.00

Figure 4: Summary Statistics Calculation

```
## Means:
V1
1    98.121340
2    51.077883
3    60.259487
dtype: float64

## Standard deviations:
          V2        V3        V4        V5         V6        V7        V8
V1
1   0.457635  0.687396  0.227141  2.539198  10.136128  0.338920  0.397361
2   0.534162  1.008391  0.313238  3.326097  16.635097  0.541507  0.700713
3   0.524689  1.076514  0.182756  2.234515  10.776433  0.353233  0.290431

          V9       V10       V11       V12       V13         V14
V1
1   0.070037  0.408849  1.238484  0.116446  0.342512  221.418938
2   0.123085  0.597813  0.918393  0.201503  0.493064  156.100173
3   0.122840  0.404555  2.286743  0.113243  0.269262  113.891805
```

Figure 5: Means and Standard deviations for each group

5 Correlation and Covariance Calculation

Pearson Correlation Coefficient: The code calculates the Pearson correlation coefficient between pairs of variables to measure the strength and direction of their linear relationship.

```
p-value:         0.18556400432462777
cor:             0.09996297573855276
           V2        V3        V4        V5        V6        V7        V8
V2   1.000000  0.099963  0.210964 -0.303350  0.258742  0.284543  0.230133
V3   0.099963  1.000000  0.164955  0.286148 -0.049049 -0.333512 -0.409324
V4   0.210964  0.164955  1.000000  0.446698  0.287107  0.128176  0.114084
V5  -0.303350  0.286148  0.446698  1.000000 -0.071707 -0.317583 -0.346922
V6   0.258742 -0.049049  0.287107 -0.071707  1.000000  0.208200  0.187101
V7   0.284543 -0.333512  0.128176 -0.317583  0.208200  1.000000  0.864046
V8   0.230133 -0.409324  0.114084 -0.346922  0.187101  0.864046  1.000000
V9  -0.151445  0.291501  0.187354  0.359395 -0.252091 -0.448301 -0.536326
V10  0.127561 -0.217975  0.008082 -0.190779  0.226504  0.610533  0.650254
V11  0.547883  0.250053  0.258643  0.020478  0.199337 -0.056401 -0.174411
V12 -0.075375 -0.560854 -0.075181 -0.272719  0.052042  0.432987  0.543208
V13  0.057417 -0.366720  0.001503 -0.268186  0.046961  0.699566  0.786372
V14  0.641068 -0.189512  0.222979 -0.436858  0.387542  0.495839  0.491180

           V9       V10       V11       V12       V13       V14
V2  -0.151445  0.127561  0.547883 -0.075375  0.057417  0.641068
V3   0.291501 -0.217975  0.250053 -0.560854 -0.366720 -0.189512
V4   0.187354  0.008082  0.258643 -0.075181  0.001503  0.222979
V5   0.359395 -0.190779  0.020478 -0.272719 -0.268186 -0.436858
V6  -0.252091  0.226504  0.199337  0.052042  0.046961  0.387542
V7  -0.448301  0.610533 -0.056401  0.432987  0.699566  0.495839
V8  -0.536326  0.650254 -0.174411  0.543208  0.786372  0.491180
...
V11  0.140192 -0.027112  1.000000 -0.522615 -0.435744  0.315632
V12 -0.261709  0.294397 -0.522615  1.000000  0.567395  0.234879
V13 -0.501859  0.513415 -0.435744  0.567395  1.000000  0.306031
V14 -0.308886  0.325731  0.315632  0.234879  0.306031  1.000000
```

Figure 6: Pearson correlation coefficient

Covariance: Covariance between variables is computed to quantify the degree to which two variables change together.

6 Principal Component Analysis (PCA)

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving most of the variance in the data.

```
Importance of components:
                    sdev               varprop              cumprop
     Standard deviation Proportion of Variance Cumulative Proportion
PC1            2.162822               0.359831              0.359831
PC2            1.581571               0.192413              0.552244
PC3            1.205541               0.111795              0.664038
PC4            0.961480               0.071111              0.735149
PC5            0.928298               0.066287              0.801437
PC6            0.803024               0.049604              0.851040
PC7            0.742955               0.042460              0.893500
PC8            0.592232               0.026980              0.920480
PC9            0.537755               0.022245              0.942725
PC10           0.496798               0.018985              0.961710
PC11           0.474805               0.017342              0.979052
PC12           0.410337               0.012952              0.992004
PC13           0.322412               0.007996              1.000000
      Standard deviation
PC1             2.162822
PC2             1.581571
PC3             1.205541
PC4             0.961480
PC5             0.928298
PC6             0.803024
PC7             0.742955
PC8             0.592232
...
PC12            0.410337
PC13            0.322412
Standard deviation    13.0
dtype: float64
```

Figure 7: Principal Component Analysis

Standardization: Before applying PCA, the data is standardized to have a mean of 0 and a standard deviation of 1, ensuring that all variables contribute equally to the analysis.

```
Mean:
             V2            V3            V4            V5            V6  \
0 -5.218675e-16  2.810056e-16 -3.813647e-16 -2.408619e-16 -8.028731e-17

             V7            V8            V9           V10           V11  \
0 -2.810056e-16  1.605746e-16 -6.021549e-16 -4.014366e-17  1.806465e-16

           V12           V13           V14
0  6.021549e-16  7.225858e-16  1.605746e-16
Standard Deviation:
    V2   V3   V4   V5   V6   V7   V8   V9  V10  V11  V12  V13  V14
0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
```

Figure 8: Standardization