Submission - Exercise [3]

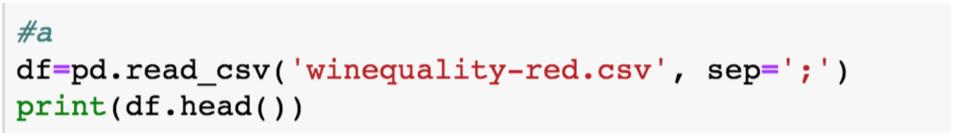
Visual Data Analysis

[Maryam Assaedi, maryam.assaedi@rwth-aachen.de]
[Mst. Mahfuja Akter , s6msakte@uni-bonn.de]
[Mahpara Hyder Chowdhury, s6machow@uni-bonn.de]

April 21, 2019

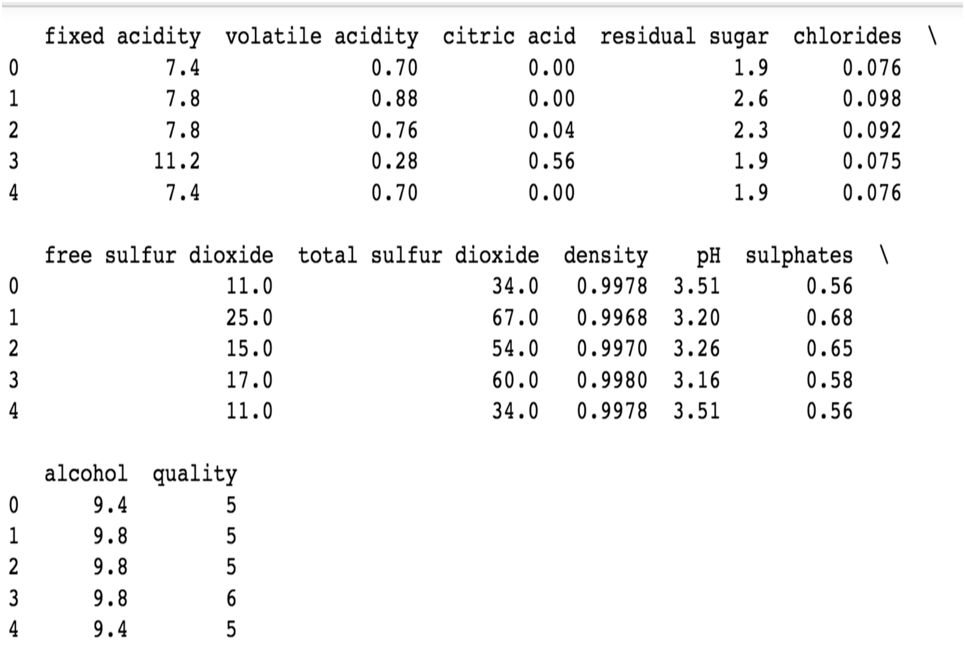
Exercise 1 (Multidimensional Data Filtering and Visualization)

a) Read the data given in winequality-red.csv, available from the lecture webpage, and print the first few rows.
We read the data and printed the first few rows using the following command:



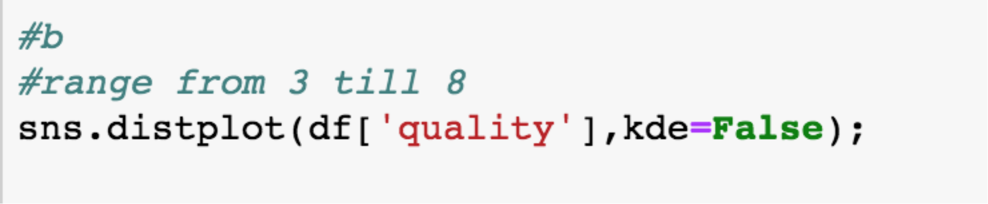
```
#a
df=pd.read_csv('winequality-red.csv', sep=';')
print(df.head())
```

which printed the following:



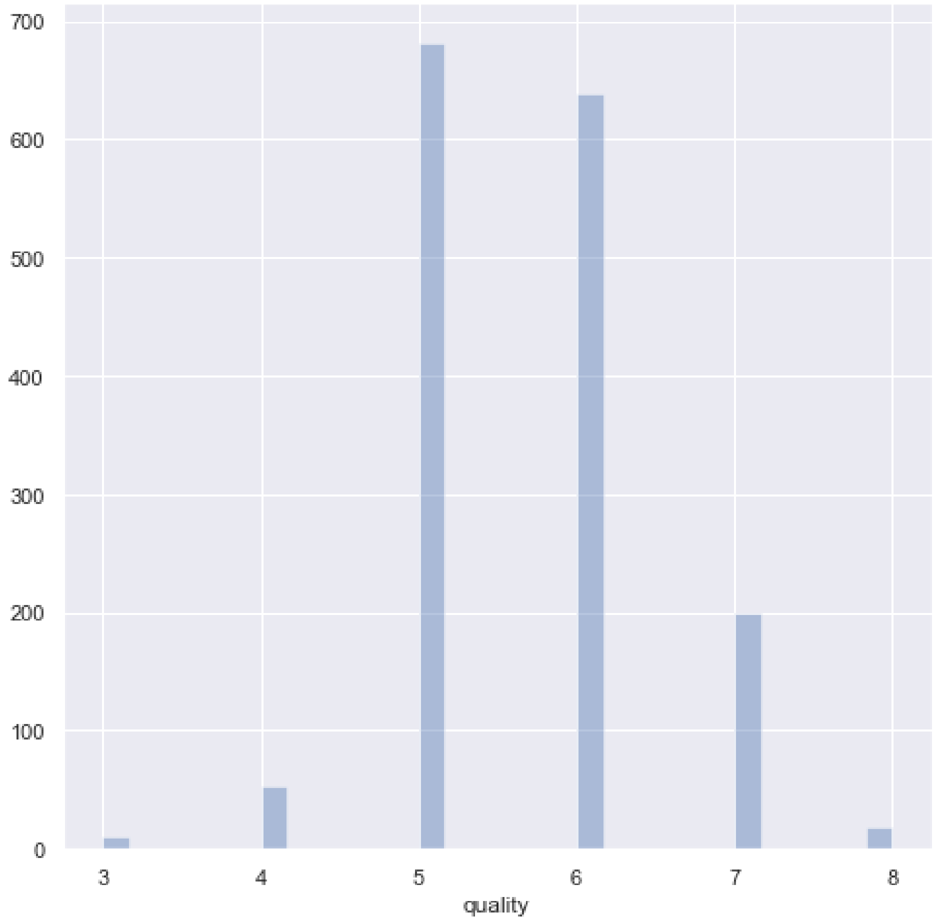
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

b) A numerical rating of sensory wine quality is given in the column "quality". Display the distribution of these scores with a histogram. What is the range of this score in the data?
To print out the histogram we used this code:



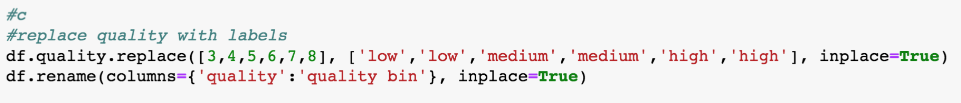
```
#b
#range from 3 till 8
sns.distplot(df['quality'],kde=False);
```

And the resulting histogram was as follows:



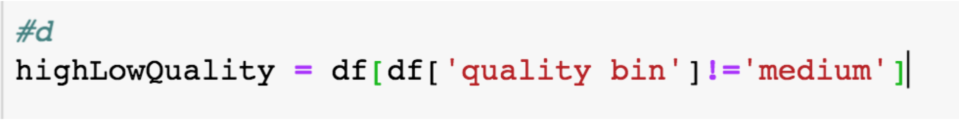
The values of the "quality" attributes range from 3 to 8.
c) Derive a coarser classification of quality into "low", "medium", and "high", by grouping together the two lowest, the intermediate, and the two highest quality scores that occur in the dataset, respectively. Replace the original "quality"

column with a new column ”quality bin” that contains these labels.
Since we have 6 values for the ”quality” attribute, we grouped scores 3&4 to ”low”, 5&6 to ”medium” and 7&8 to ”high”. Then we changed the attribute’s name to ”quality bin”.



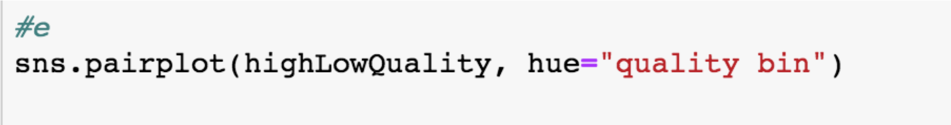
```
#c
#replace quality with labels
df.quality.replace([3,4,5,6,7,8], ['low','low','medium','medium','high','high'], inplace=True)
df.rename(columns={'quality':'quality bin'}, inplace=True)
```

d) We would like to investigate differences between high and low quality wines. Therefore, create a filtered data frame in which the medium-quality wines are omitted.
We created a new dataframe and called it ”highLowQuality”



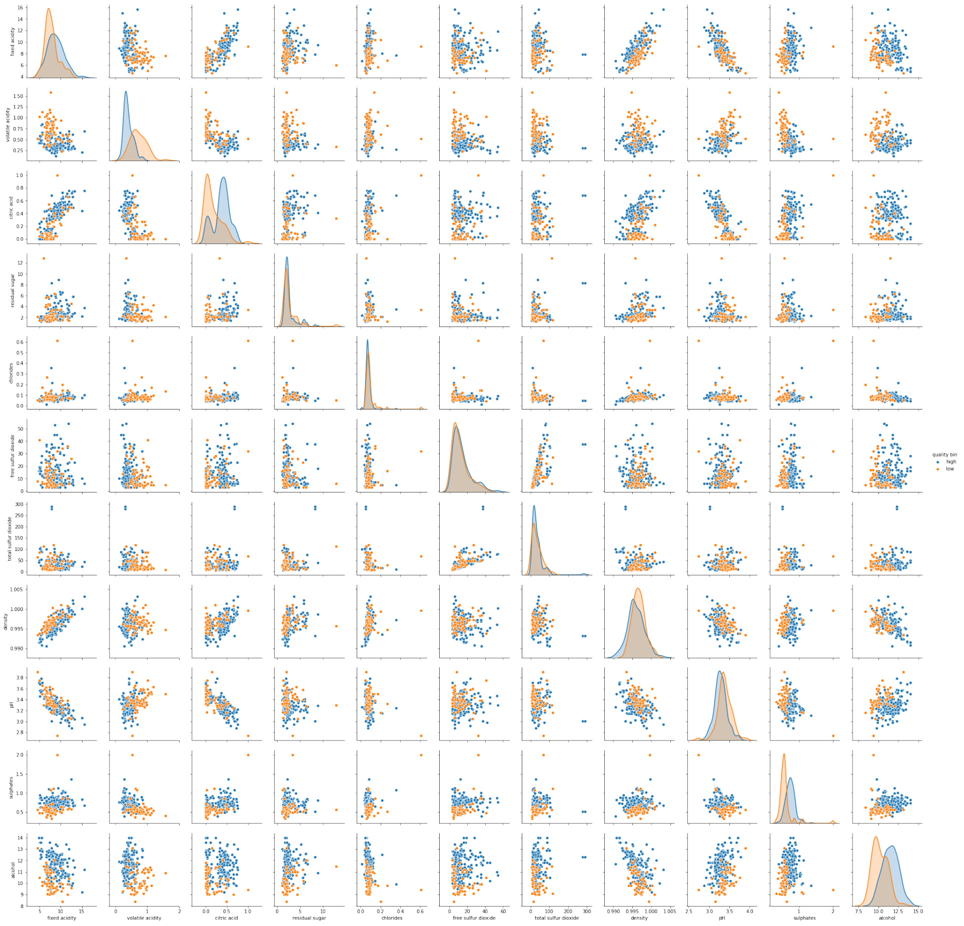
```
#d
highLowQuality = df[df['quality bin']!='medium']
```

e) Visualize all numerical attributes in a scatterplot matrix. Color the two quality levels differently.
We used the PairGrid class



```
#e
sns.pairplot(highLowQuality, hue="quality bin")
```

to print out the plot

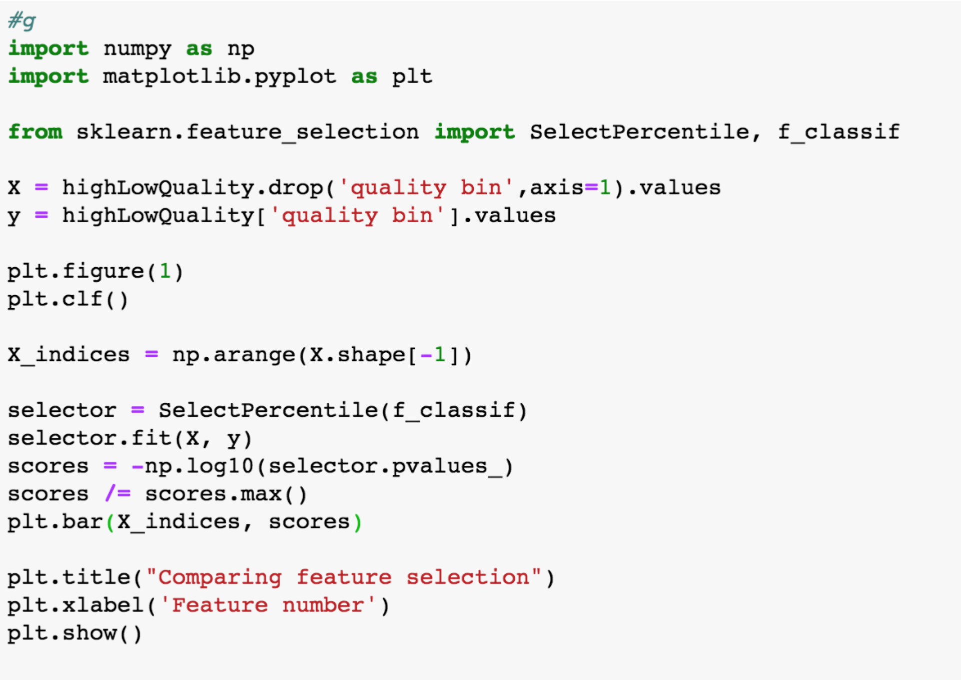


f) Based on the visualization, name five attributes that appear to best distinguish between high and low quality.

From looking at the diagonal density estimates, we can search for attributes that have a range of values that is different for each quality class. We can notice a clear distinction in the ”volatile acidity”, ”citric acid”,”sulphates” and ”alcohol” attributes. For the fifth contributing feature, we considered ”pH” and ”density”. But because ”pH” has a higher displacement than ”density” we think it contributes more to the quality.

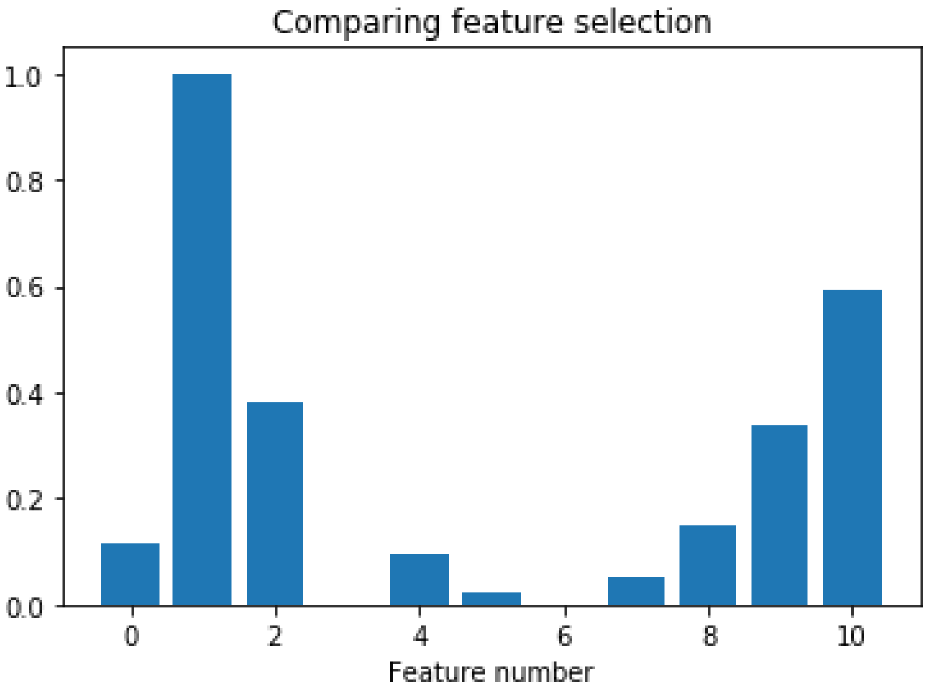
g) Now, use an automated feature selection technique to identify five attributes that distinguish between high and low quality. More specifically, please use the F score from a one-way analysis of variance (ANOVA) to rank the attributes, as implemented in scikit learn’s f_classif. What are the five best attributes according to this measure? Are they the same as those you identified visually? Create a filtered data frame that only contains the top five attributes, plus the ”quality bin”.

The code we used to automatically select features is as follows:



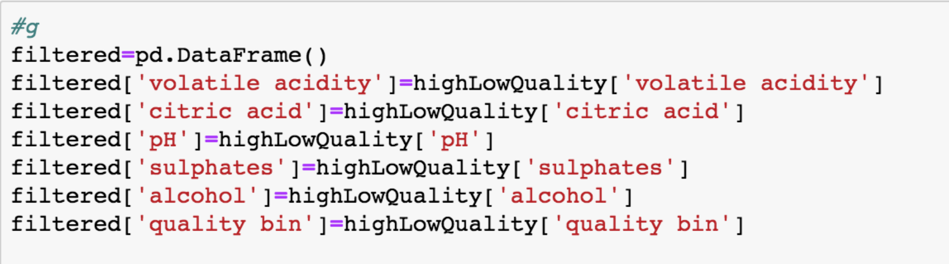
```
#g
import numpy as np
import matplotlib.pyplot as plt

from sklearn.feature_selection import SelectPercentile, f_classif

X = highLowQuality.drop('quality bin',axis=1).values
y = highLowQuality['quality bin'].values

plt.figure(1)
plt.clf()

X_indices = np.arange(X.shape[-1])

selector = SelectPercentile(f_classif)
selector.fit(X, y)
scores = -np.log10(selector.pvalues_)
scores /= scores.max()
plt.bar(X_indices, scores)

plt.title("Comparing feature selection")
plt.xlabel('Feature number')
plt.show()
```

The output plot is as follows



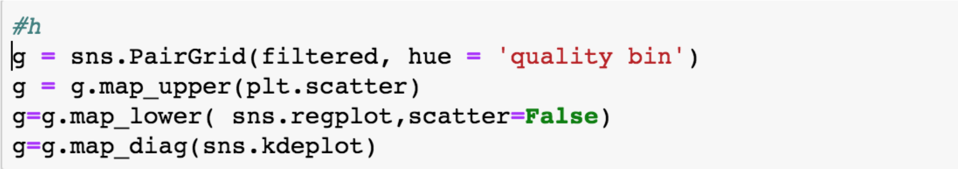
From the plot we can see that, as expected, the ”volatile acidity”, ”citric acid”,”sulphates” and ”alcohol” have high contribution to the ”quality” attribute. Even our conclusion about the ”pH” attribute turned out to be correct.

Finally, we created a new dataFrame called ”filtered” with the 5 contributing attributes plus ”quality bin” as follows:



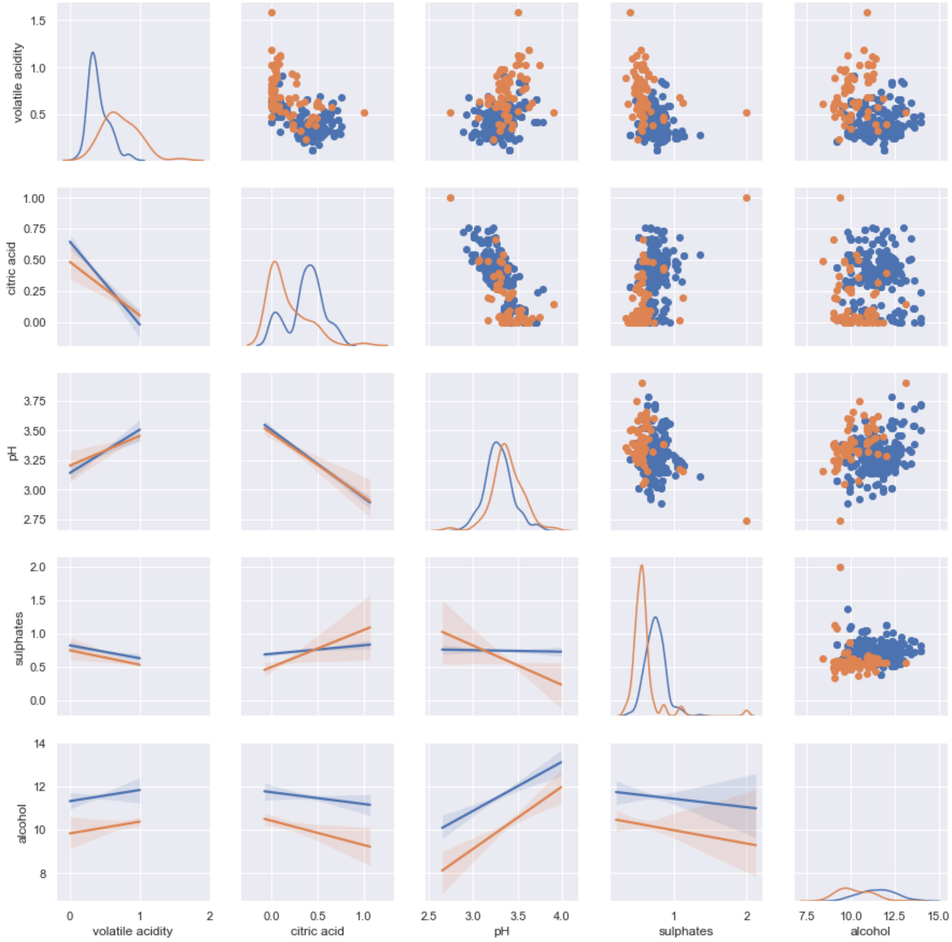
```
#g
filtered=pd.DataFrame()
filtered['volatile acidity']=highLowQuality['volatile acidity']
filtered['citric acid']=highLowQuality['citric acid']
filtered['pH']=highLowQuality['pH']
filtered['sulphates']=highLowQuality['sulphates']
filtered['alcohol']=highLowQuality['alcohol']
filtered['quality bin']=highLowQuality['quality bin']
```

h) Create a matrix similar to the one in Fig. 1: It should compare the two quality bins with respect to the five top-ranking attributes, using density estimates (on the diagonal), scatterplots (in the upper triangular part), and plots of pairwise linear regression models (in the lower triangular part).
We managed to plot the required plot with the following code:



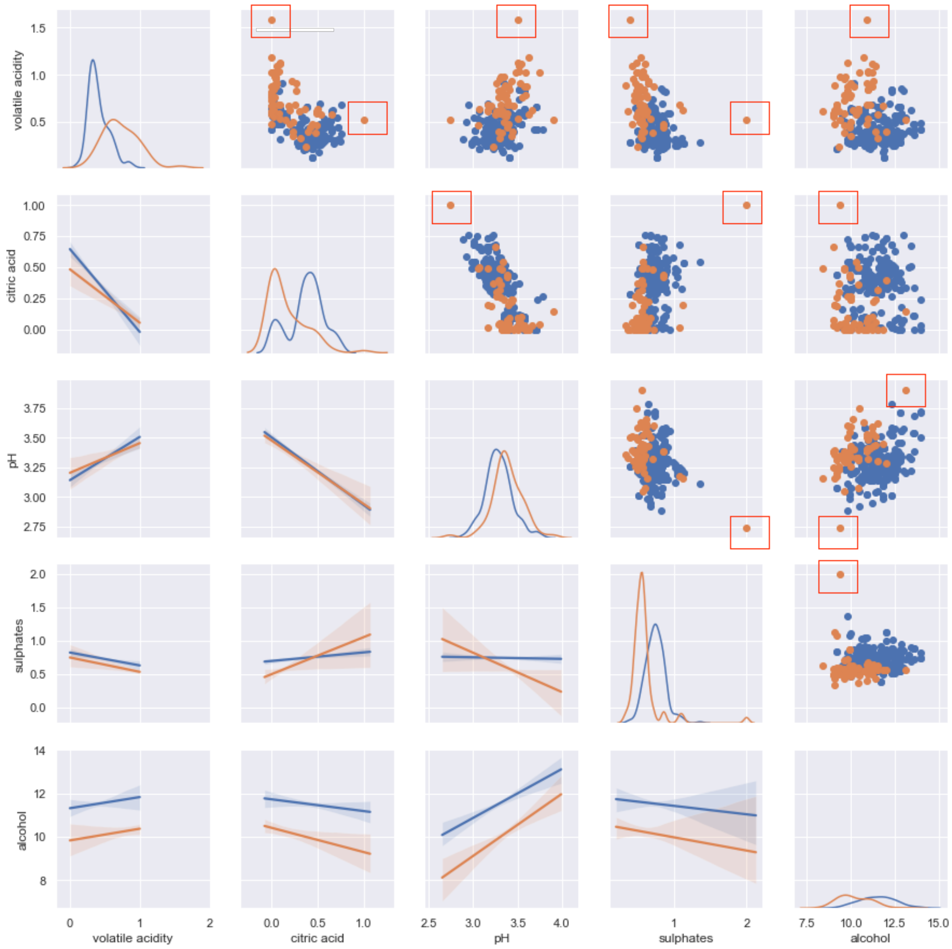
```
#h
g = sns.PairGrid(filtered, hue = 'quality bin')
g = g.map_upper(plt.scatter)
g=g.map_lower( sns.regplot,scatter=False)
g=g.map_diag(sns.kdeplot)
```

To get the following plot



i) Based on the visualization, which attributes appear to be strongly correlated regardless of quality? For which attributes does the amount of correlation appear to depend on the quality? Does any of the attributes appear to have a multimodal distribution? Point out one or multiple data points that appear to be outliers?

Attributes that seem highly correlated regardless of quality: citric acid with pH
Attributes that seem highly correlated with dependence on quality: alcohol with all other attributes and sulphates with volitile acidity
Attributes with multimodal distribution : citric acid Outliers have been marked with e red square in the following plot



Exercise 2 (Evaluating PCP Variants)

a) In addition to standard PCPs, the authors test eight variations. For each of them, briefly describe the proposed modification, and why the authors expected it to improve the visualization. Use 1-2 sentences for each of the eight cases.
Answer:
To facilitate multivariate data correlation and cluster identification, reduce visual clutter and increase information throughput, authors analyzed cluster identification performance of nine PCP variations, including standard PCPs. Those variations are given below:
ColorBlend: When rendering semitransparent objects, this method have significant impact on perception of order and structure. Depth ordering clue is better represented in this method.
Color : Color has a very strong visual cue and thus expected to get positive reinforcement from multiple cues.
SP: Scatter plot has intuitive scatter plot. The data are displayed as a collection of points, thus have a better visual cue.
Curved: It could resolves data ambiguities and help to following poly-lines across

axis.
Blend: It has stronger cue than parallax because of noise suppression.
Wobble: It can generates motion parallax but keeps cluster fixed and eases visual tracking.
RT(Random tour) and PT(Permutation tour): Another way to facilitate cluster identification is animation to PCPs. RT and PT inspires the animation schemes which helps to visually separate clusters.
All above variance to outperform standard due to the presence of additional cues. For each variance, response time measurements were normalized per cluster count which is being used as input of ANOVA and post-hoc analysis.

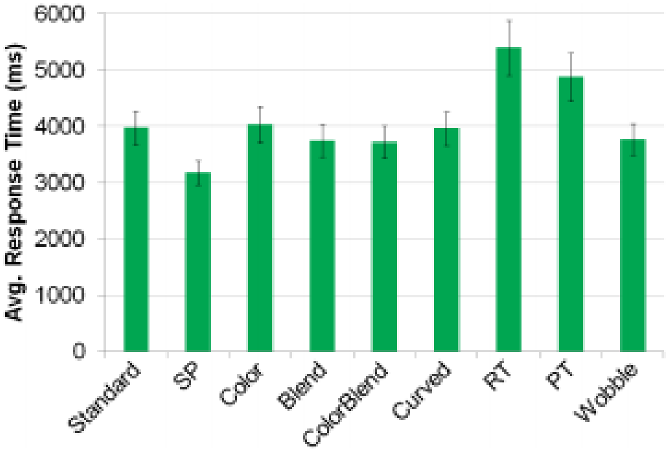


Figure 8: *Average response time for each of the PCP variations (shown with a 95% confidence interval).*

From above graph, we see that only SP performs significantly better than eight remaining variations. RT and PT perform even worse than standard PCP. SPs were mainly chosen because they are one of the most multivariate visualization technique and limit the number of PCP variations to keep testing feasible.
b) Within the user study, what task did the subjects have to perform? Name another task for which Parallel Coordinates are frequently used in practice, but which was not included in the study.
Answer:
Within User study, each participants was asked to perform a number of cluster identification trials. For each trials they had to answer the number of cluster, response time and correctness were being displayed. Based on this, authors generated various stimuli of between two to six clusters.
Parallel coordinates were frequently used in the identification of correlation between data variables in high dimensional data sets (2D and 3D PCPs) using both 2D scatter plots and 2D PCPs.
c) To which extent did the results of the study match the authors' hypotheses?
Answer:
With respect to increasing cluster count, both response time and correctness, authors hypothesis and user studies gives similar kind of increase and decrease patterns for all PCP variants.

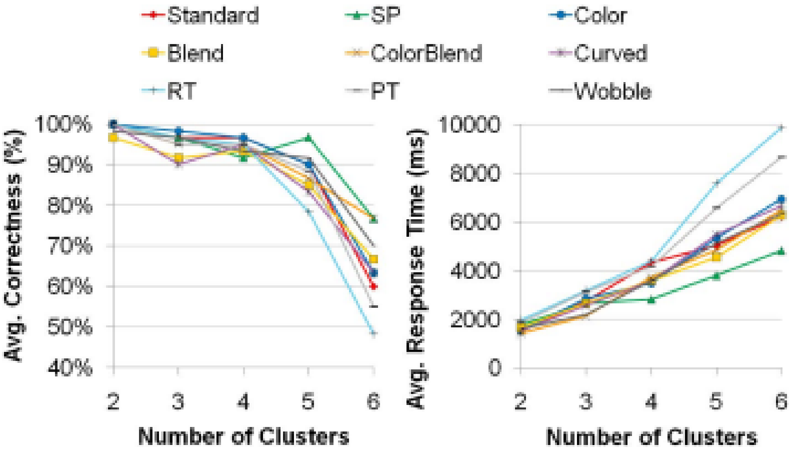


Figure 6: *Response time and correctness plotted against cluster count show similar patterns for all PCP variations.*

d) Which of the explored modifications would you consider using when designing a visualization based on Parallel Coordinates? Briefly justify your answer.
Answer:
Designing a multivariate data for visualization based on PCP, The scatter plot is the best one. SP contained visualization elements that were non-PCP variant. Scatter plots are more effective than PCPs in supporting visual correlation analysis. Identifying the number of clusters, a PCP representation might still be better suited for certain qualitative analysis (ex: obtaining insight in the actual shapes of cluster). We can explore this method for better visualization cue to limit overdraw problem.

Exercise 3 (Multidimensional Data Visualization)

a) Does the ordering of attributes have a greater effect on a parallel coordinates plot, or on a scatter-plot matrix? Briefly explain why
Answer:
Both scatter-plot matrix and parallel coordinates plot uses ordering technique of multivariate data. But in parallel coordinates the ordering have greater effect, because k equidistant axes are parallel to one of screen axes (which is ordered) and correspond to the attributes. And every data item corresponds to a polygonal line which intersects each of axis to corresponding value of attribute.
b) Briefly explain a limitation of traditional scatterplots and a way to overcome it.
Answer:
The limitations of traditional scatterplot is overdrawing of values. Translucency is a powerful tool for dealing with over plotting.
c) Briefly describe the difference between parallel coordinates and star coordinates. Which of the two would you select to explore correlations between

values along two specific axes? How would you be able to detect a correlation in the type of plot you have chosen?

Answer:

Parallel Coordinates each data element is represented as a line passing through the coordinate axes. And Star Coordinates arranges coordinates on a circle sharing the same origin at the center which points to represent data, treating each dimension uniformly at the cost of coarse representation.

Human vision can effectively explore certain data structure for D at least up to 100(according to researcher). This is contrary MDA such as Parallel Coordinates. They are based on the idea that human vision can not directly see data structures in the D¿3 data space but can analyze and describe data structures indirectly if the data attributes are mapped into visible image features. Therefore better to use this capability of stars before rejection of the most part of unknown.

It has been found to be particularly useful in gaining insight into hierarchically clustered datasets.

d) In Chapter 3, Slide 34, we claimed that, when we take points on a line in Cartesian coordinates, and display them in parallel coordinates, they will meet in a common point. Perform a computation based on the equations on that slide to verify this claim

Answer:

Given that,

Cartesian coordinates (x1,x2) represented by parallel coordinates line by

y = (x2-x1)*x + x1 ——————(1)

and if m !=1 then parallel coordinates (1/(1-m) , b/(1-m)) points intersects with the line of Cartesian coordinates x2 = m*x1 + b.

We put the intersecting points value in x,y coordinates in equation 1.

b/(1-m) = (x2-x1)* 1/(1-m) + x1

b = (x2-x1) + x1*(1-m) [multiplying (1-m) in both side]

b = x2 - x1 + x1 - m*x1

b = x2 - m*x1

x2 = m*x1 + b

which is equivalent to line of Cartesian coordinates.

we can say that both Cartesian and parallel coordinates lines intersect in given points.