Wine_visualisation

January 16, 2025

1 Wine Data Visualisation in Jupyter Notebook - using atoti

Link to the tutorial: https://www.youtube.com/watch?v=Y49662c3EL4&ab_channel=LightsOnData - **Youtube: How to Create a Data Visualization in Jupyter Notebook Using atoti**

[2]:

```
import atoti as tt
import numpy as np
import pandas as pd
import seaborn as sns
```

2 Read the data for the red and white wines

2.1 Red Wine

[6]:

```
wine_red = pd.read_csv(
    "https://data.atoti.io/notebooks/wine-analytics/winequality-red.csv", sep=";
  ↪"
)
wine_red.head()
```

[6]:

```
   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
```

2.2 White Wine

[7]:

```
wine_white = pd.read_csv(
    "https://data.atoti.io/notebooks/wine-analytics/winequality-white.csv",␣
  ↪sep=";"
)
wine_white.head()
```

[7]:

```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.0              0.27         0.36            20.7      0.045
1            6.3              0.30         0.34             1.6      0.049
2            8.1              0.28         0.40             6.9      0.050
3            7.2              0.23         0.32             8.5      0.058
4            7.2              0.23         0.32             8.5      0.058

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 45.0                 170.0   1.0010  3.00       0.45
1                 14.0                 132.0   0.9940  3.30       0.49
2                 30.0                  97.0   0.9951  3.26       0.44
3                 47.0                 186.0   0.9956  3.19       0.40
4                 47.0                 186.0   0.9956  3.19       0.40

   alcohol  quality
0      8.8        6
1      9.5        6
2     10.1        6
3      9.9        6
4      9.9        6
```

2.3 Add a new column “category” to the datasets

[9]:

```
wine_red["category"] = "Red"
wine_red.head()
```

[9]:

```
   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 category
0      9.4        5      Red
1      9.8        5      Red
2      9.8        5      Red
3      9.8        6      Red
4      9.4        5      Red
```

[11]:

```
wine_white["category"] = "White"
wine_white.head()
```

[11]:

```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.0              0.27         0.36            20.7      0.045
1            6.3              0.30         0.34             1.6      0.049
2            8.1              0.28         0.40             6.9      0.050
3            7.2              0.23         0.32             8.5      0.058
4            7.2              0.23         0.32             8.5      0.058

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 45.0                 170.0   1.0010  3.00       0.45
1                 14.0                 132.0   0.9940  3.30       0.49
2                 30.0                  97.0   0.9951  3.26       0.44
3                 47.0                 186.0   0.9956  3.19       0.40
4                 47.0                 186.0   0.9956  3.19       0.40

   alcohol  quality category
0      8.8        6    White
1      9.5        6    White
2     10.1        6    White
3      9.9        6    White
4      9.9        6    White
```

[12]:

```
wines = pd.concat([wine_red, wine_white], axis=0, ignore_index=True)
wines.index.set_names("wine index", inplace=True)
wines.head()
```

[12]:

```
            fixed acidity  volatile acidity  citric acid  residual sugar  \
wine index
0                     7.4              0.70         0.00             1.9
1                     7.8              0.88         0.00             2.6
2                     7.8              0.76         0.04             2.3
3                    11.2              0.28         0.56             1.9
4                     7.4              0.70         0.00             1.9
```

```
            chlorides  free sulfur dioxide  total sulfur dioxide  density  \
wine index
0               0.076                 11.0                  34.0   0.9978
1               0.098                 25.0                  67.0   0.9968
2               0.092                 15.0                  54.0   0.9970
3               0.075                 17.0                  60.0   0.9980
4               0.076                 11.0                  34.0   0.9978

              pH  sulphates  alcohol  quality category
wine index
0           3.51       0.56      9.4        5      Red
1           3.20       0.68      9.8        5      Red
2           3.26       0.65      9.8        5      Red
3           3.16       0.58      9.8        6      Red
4           3.51       0.56      9.4        5      Red
```

[13]:

```
wines["alcohol range"] = wines["alcohol"].apply(np.floor)
```

[14]:

```
wines["Rating"] = "Good"
wines.loc[wines["quality"] < 7, "Rating"] = "Average"
wines.loc[wines["quality"] < 5, "Rating"] = "Poor"
wines
```

[14]:

```
            fixed acidity  volatile acidity  citric acid  residual sugar  \
wine index
0                     7.4              0.70         0.00             1.9
1                     7.8              0.88         0.00             2.6
2                     7.8              0.76         0.04             2.3
3                    11.2              0.28         0.56             1.9
4                     7.4              0.70         0.00             1.9
…                       …               …          …             …
6492                  6.2              0.21         0.29             1.6
6493                  6.6              0.32         0.36             8.0
6494                  6.5              0.24         0.19             1.2
6495                  5.5              0.29         0.30             1.1
6496                  6.0              0.21         0.38             0.8

            chlorides  free sulfur dioxide  total sulfur dioxide  density  \
wine index
0               0.076                 11.0                  34.0  0.99780
1               0.098                 25.0                  67.0  0.99680
2               0.092                 15.0                  54.0  0.99700
3               0.075                 17.0                  60.0  0.99800
4               0.076                 11.0                  34.0  0.99780
…                 …                  …                   …      …
6492            0.039                 24.0                  92.0  0.99114
6493            0.047                 57.0                 168.0  0.99490
```

```
6494            0.041                 30.0                 111.0  0.99254
6495            0.022                 20.0                 110.0  0.98869
6496            0.020                 22.0                  98.0  0.98941

              pH  sulphates  alcohol  quality category  alcohol range   Rating
wine index
0           3.51       0.56      9.4        5      Red            9.0  Average
1           3.20       0.68      9.8        5      Red            9.0  Average
2           3.26       0.65      9.8        5      Red            9.0  Average
3           3.16       0.58      9.8        6      Red            9.0  Average
4           3.51       0.56      9.4        5      Red            9.0  Average
…          …        …      …      …      …            …      …
6492        3.27       0.50     11.2        6    White           11.0  Average
6493        3.15       0.46      9.6        5    White            9.0  Average
6494        2.99       0.46      9.4        6    White            9.0  Average
6495        3.34       0.38     12.8        7    White           12.0     Good
6496        3.26       0.32     11.8        6    White           11.0  Average

[6497 rows x 15 columns]
```

[25]:

[25]:

```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0               7.4             0.700         0.00             1.9       0.076
1               7.8             0.880         0.00             2.6       0.098
2               7.8             0.760         0.04             2.3       0.092
3              11.2             0.280         0.56             1.9       0.075
4               7.4             0.700         0.00             1.9       0.076
…             …               …          …             …        …
1594            6.2             0.600         0.08             2.0       0.090
1595            5.9             0.550         0.10             2.2       0.062
1596            6.3             0.510         0.13             2.3       0.076
1597            5.9             0.645         0.12             2.0       0.075
1598            6.0             0.310         0.47             3.6       0.067

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                    11.0                  34.0  0.99780  3.51       0.56
1                    25.0                  67.0  0.99680  3.20       0.68
2                    15.0                  54.0  0.99700  3.26       0.65
3                    17.0                  60.0  0.99800  3.16       0.58
4                    11.0                  34.0  0.99780  3.51       0.56
…                   …                   …      …   …        …
1594                 32.0                  44.0  0.99490  3.45       0.58
1595                 39.0                  51.0  0.99512  3.52       0.76
1596                 29.0                  40.0  0.99574  3.42       0.75
1597                 32.0                  44.0  0.99547  3.57       0.71
1598                 18.0                  42.0  0.99549  3.39       0.66
```

```
      alcohol  quality
0         9.4        5
1         9.8        5
2         9.8        5
3         9.8        6
4         9.4        5
…        …       …
1594     10.5        5
1595     11.2        6
1596     11.0        6
1597     10.2        5
1598     11.0        6

[1599 rows x 12 columns]
```

2.3.1 Prepare the data for corelation

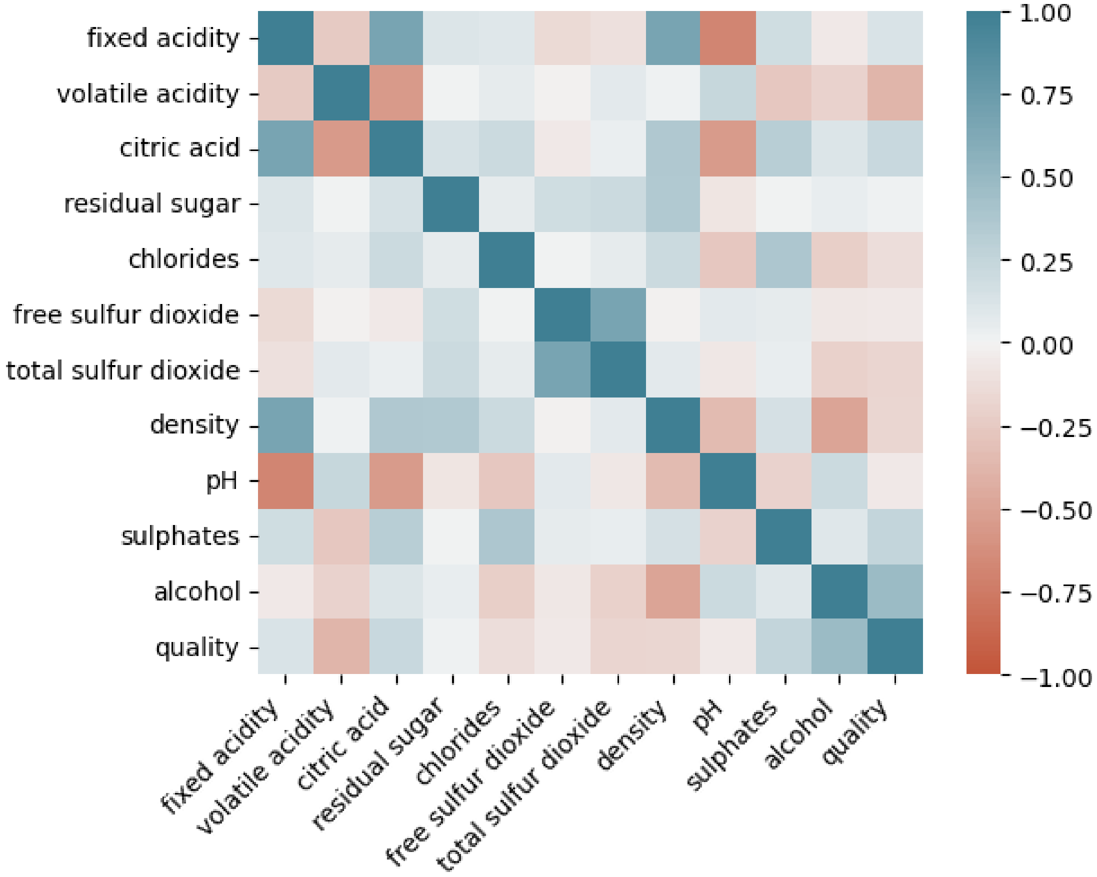
[27]:

```
onlyRedWines = wine_red[wine_red['category'] == 'Red']
redWinesReadyForCorelation = onlyRedWines.drop(columns=['category'])
```

2.4 Default corelation => using Pearson method

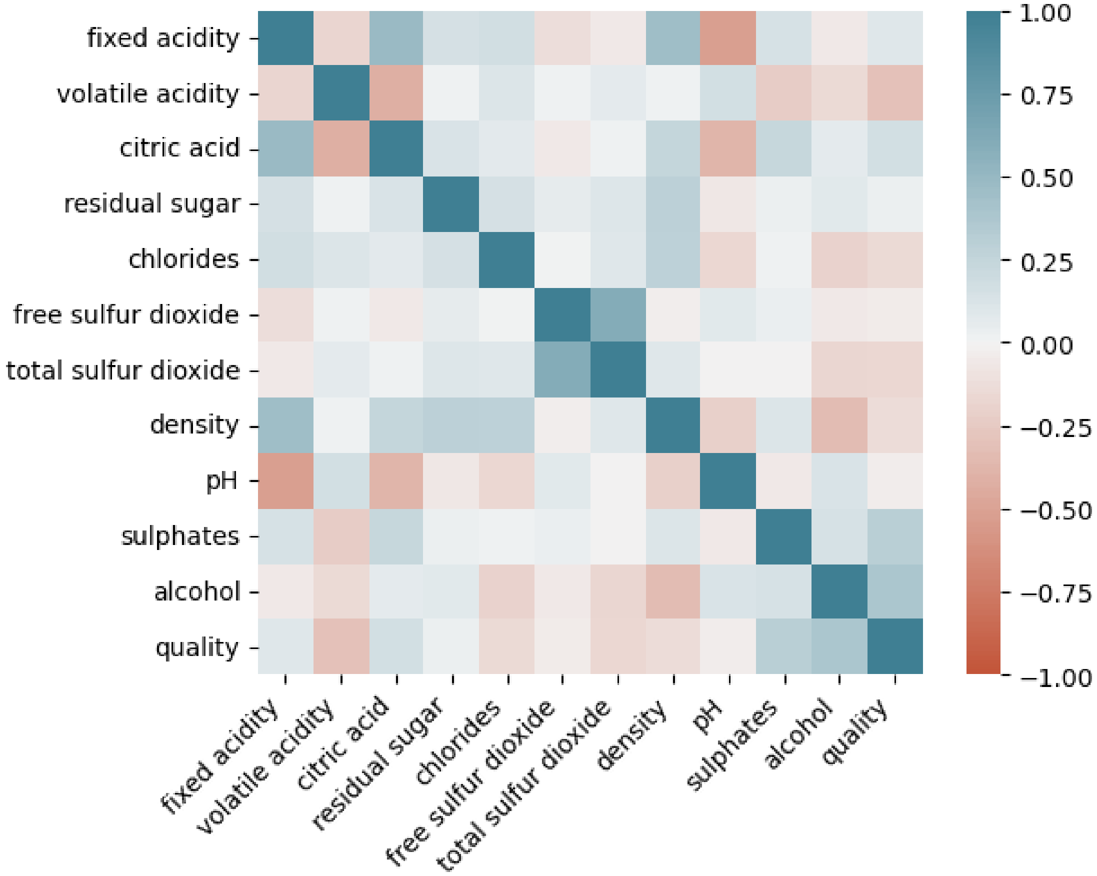
[28]:

```
corr_red_pearson = redWinesReadyForCorelation.corr()
ax_red = sns.heatmap(
    corr_red_pearson,
    vmin=-1,
    vmax=1,
    center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True,
)
ax_red.set_xticklabels(
    ax_red.get_xticklabels(), rotation=45, horizontalalignment="right"
);
```



2.5 Corelation using kendall method

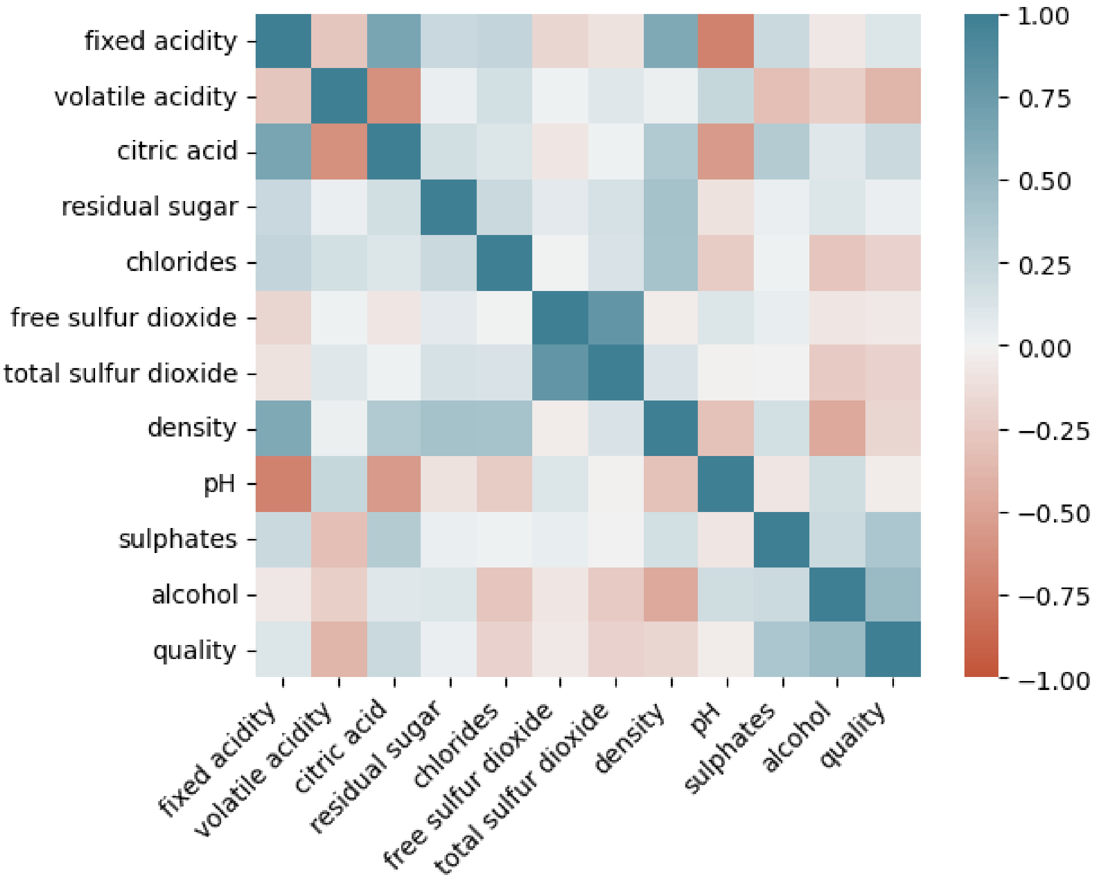
```
[29]: corr_red_kendall = redWinesReadyForCorelation.corr(method='kendall')
ax_red = sns.heatmap(
    corr_red_kendall,
    vmin=-1,
    vmax=1,
    center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True,
)
ax_red.set_xticklabels(
    ax_red.get_xticklabels(), rotation=45, horizontalalignment="right"
);
```



2.6 Corelation using spearman method

[30]:

```
corr_red_spearman = redWinesReadyForCorelation.corr(method='spearman')
ax_red = sns.heatmap(
    corr_red_spearman,
    vmin=-1,
    vmax=1,
    center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True,
)
ax_red.set_xticklabels(
    ax_red.get_xticklabels(), rotation=45, horizontalalignment="right"
);
```



```
[40]: print("Pearson corelation values")
      corr_red_pearson[(corr_red_pearson['quality'] > 0.25) |␣
        ↪(corr_red_pearson['quality'] < -0.25)]["quality"]
```

```
Pearson corelation values
```

```
[40]: volatile acidity   -0.390558
      sulphates           0.251397
      alcohol             0.476166
      quality             1.000000
      Name: quality, dtype: float64
```

```
[42]: print("Kendall corelation values")
      corr_red_kendall[(corr_red_kendall['quality'] > 0.25) |␣
        ↪(corr_red_kendall['quality'] < -0.25)]["quality"]
```

```
Kendall corelation values
```

```
[42]: volatile acidity   -0.300779
      sulphates           0.299270
```

```
alcohol             0.380367
quality             1.000000
Name: quality, dtype: float64
```

```
[43]: print("Spearman corelation values")
      corr_red_spearman[(corr_red_spearman['quality'] > 0.25) |␣
       ↪(corr_red_spearman['quality'] < -0.25)]["quality"]
```

```
Spearman corelation values
```

```
[43]: volatile acidity   -0.380647
      sulphates           0.377060
      alcohol             0.478532
      quality             1.000000
      Name: quality, dtype: float64
```

Pearson - Linear relashiopnship - continuous variables - normally distributed data - homoscedasticity - the variability is the same between dependent and non dependent variables

Spearman - Monotonic relashionship - continous or ordinal - non-parametric => does not assume data's distribution - variables are ranked => a ranking has to be done before the analasys is done

BOTH: - look at linear relations (they only decrease or increase...) -

```
[ ]:
```

```
[ ]: b
```