Capstone-Projekt: Eine Analyse zur Weinqualität

Analyse der Weinqualität mit Python, Power BI und Excel

Titel: Eine Analyse zur Weinqualität
Name: Ahmadi,Seyed Mohammad Hossein
Datum: 1 April 2025

Einleitung

Ziel des Projekts und Verwendete Tools

Ziel des Projekts:

- Analyse der Faktoren, die die Weinqualität beeinflussen.
- Vorhersage der Weinqualität mit verschiedenen Tools (Python, Power BI, Excel).

Verwendete Tools:

- **Python:** Pandas, Matplotlib, Seaborn, Scikit-Learn für Machine Learning (Random Forest, XGBoost).
- **Power BI:** Interaktive Dashboards und Matrix Visualization für die Datenanalyse.
- **Excel:** Pivot Tables und Diagramme zur grundlegenden Analyse.

Modellvergleich:

- **Beste Ergebnisse:** Random Forest Modell mit einem R^2 von 0.463 und MSE von 0.299.

Datensatzbeschreibung

- **Datensatz:** Enthält verschiedene chemische Eigenschaften von Weinen
- **Anzahl der Einträge:** 1143
- **Anzahl der Merkmale:** 13
- **Wichtigste Features:** Alkoholgehalt, pH-Wert, Zitronensäure etc.

1. Ziel des Projekts

- Untersuchung der Faktoren, die die Weinqualität beeinflussen.
- Analyse der Merkmale wie Alkoholgehalt, pH-Wert, Säuregehalt etc.
- Vorhersage der Weinqualität mit Python und optional Power BI und Excel.

2. Verwendete Tools

- Haupttool: Python
- Pandas: Zur Datenmanipulation und -bereinigung.
- Matplotlib & Seaborn: Für die Visualisierung von Daten (z. B. Histogramme, Streudiagramme, Korrelationsmatrix).
- Scikit-Learn: Zur Implementierung und Evaluierung von Machine Learning-Modellen.
- XGBoost & Random Forest: Modelle für die Vorhersage der Weinqualität.
- Zusätzliche Tools: Power BI und Excel
- Power BI: Interaktive Dashboards und Visualisierungen.
- Excel: Pivot-Tabellen und Diagramme zur grundlegenden Analyse.

3. Datenvorbereitung und Analys (Pythonschwerpunkt)

- Daten einlesen und erste Inspektion mit Pandas.
- Statistische Zusammenfassung der Daten: Mittelwert, Standardabweichung etc.
- Visualisierung:
- Histogramme für die Verteilung der Merkmale.
- Korrelationsmatrix zur Analyse der Beziehungen zwischen den Variablen.
- Datenaufbereitung:
- Entfernung von Ausreißern.
- Normalisierung und Standardisierung der Daten.
- Aufteilung der Daten in Trainings- und Testdaten (80% - 20%).

[3]:
```
# مرحله 3: بررسی داده‌ها و اطلاعات اولیه
# Schritt 3: Daten betrachten und grundlegende Informationen erhalten
print(df.head())  # اولین 5 ردیف داده‌ها را نمایش می‌دهد
# Zeigt die ersten 5 Zeilen der Daten an
print(df.info())  # اطلاعات کلی در مورد ستون‌ها
# Zeigt grundlegende Informationen zu den Spalten an
```

```
   fixed acidity  volatile acidity  citric acid  residual sugar  chloride
0            7.4              0.70         0.00             1.9     0.07
1            7.8              0.88         0.00             2.6     0.09
2            7.8              0.76         0.04             2.3     0.09
3           11.2              0.28         0.56             1.9     0.07
4            7.4              0.70         0.00             1.9     0.07

   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  Id
0      9.4        5   0
1      9.8        5   1
2      9.8        5   2
3      9.8        6   3
4      9.4        5   4
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64
 12  Id                    1143 non-null   int64
dtypes: float64(11), int64(2)
memory usage: 116.2 KB
None
```

[11]:
```
# مرحله 5: تقسیم داده‌ها به داده‌های آموزشی و آزمایشی
# Schritt 5: Daten in Trainings- und Testdaten aufteilen
X = df.drop(['quality', 'Id'], axis=1)  # ویژگی‌ها (بدون ستون کیفیت و ID)
# Merkmale (ohne die Spalten "quality" und "Id")
y = df['quality']  # هدف (کیفیت شراب)
# Ziel (Weinqualität)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  # تقسیم داده‌ها به 80% آموزش و 20% تست
# Aufteilung der Daten in 80% Training und 20% Test
```

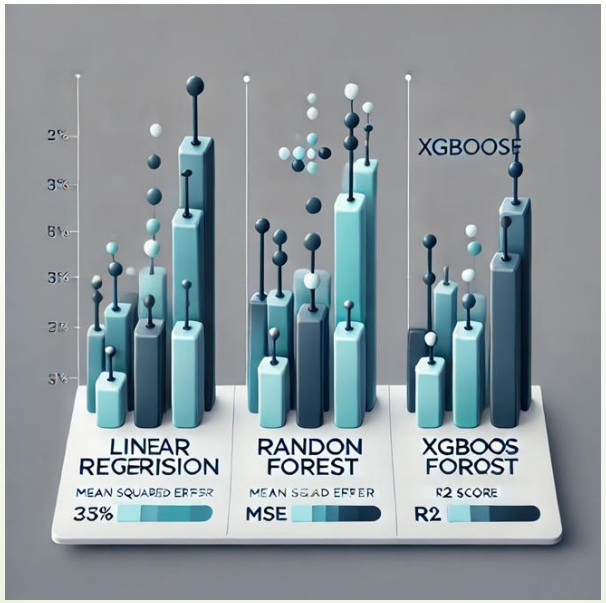
[13]:
```
# مرحله 6: ساخت مدل رگرسیون خطی
# Schritt 6: Lineares Regressionsmodell erstellen
model = LinearRegression()  # ایجاد مدل رگرسیون خطی
# Erstellen eines linearen Regressionsmodells
model.fit(X_train, y_train)  # آموزش مدل بر روی داده‌های آموزشی
# Modell mit den Trainingsdaten trainieren
```

[13]:
```
LinearRegression()
```

[15]:
```
# مرحله 7: پیش‌بینی بر روی داده‌های آزمایشی
# Schritt 7: Vorhersagen für Testdaten machen
y_pred = model.predict(X_test)  # پیش‌بینی کیفیت شراب با استفاده از مدل
# Vorhersage der Weinqualität mit dem Modell
```

4. Modellvergleich und Ergebnisse

- Vergleich der Modelle:
- Lineare Regression: MSE = 0.38, R^2 = 0.32
- Random Forest: MSE = 0.30, R^2 = 0.46
- XGBoost: MSE = 0.37, R^2 = 0.34
- Bestes Modell: Random Forest zeigte die besten Ergebnisse mit einem R^2 von 0.463 und einem MSE von 0.299.



[21]:

```
# وارد کردن کتابخانه‌های XGBoost
import xgboost as xgb  # وارد کردن مدل XGBoost  /  XGBoost-Modell importieren

# ایجاد مدل XGBoost
xgb_model = xgb.XGBRegressor(n_estimators=100, random_state=42)  # تنظیمات مدل  /  Modell konfigurieren

# آموزش مدل روی داده‌های آموزشی
xgb_model.fit(X_train, y_train)  # آموزش مدل با استفاده از داده‌های آموزشی  /  Modell mit Trainingsdaten trainieren

# پیش‌بینی با استفاده از مدل آموزش‌دیده
y_pred_xgb = xgb_model.predict(X_test)  # پیش‌بینی با استفاده از داده‌های تست  /  Vorhersagen mit Testdaten machen

# ارزیابی مدل
mse_xgb = mean_squared_error(y_test, y_pred_xgb)  # محاسبه خطای مربعات میانگین  /  Mittleren quadratischen Fehler berechnen
r2_xgb = r2_score(y_test, y_pred_xgb)  # محاسبه ضریب تعیین  /  Bestimmtheitsmaß berechnen

print(f"Mean Squared Error (XGBoost): {mse_xgb}")  # نمایش خطای مربعات میانگین  /  Mittleren quadratischen Fehler anzeigen
print(f"R² Score (XGBoost): {r2_xgb}")  # نمایش ضریب تعیین  /  Bestimmtheitsmaß anzeigen
```

```
Mean Squared Error (XGBoost): 0.36556705832481384
R² Score (XGBoost): 0.3430641293525696
```

[23]:

```
# وارد کردن کتابخانه‌های شبکه عصبی
from sklearn.neural_network import MLPRegressor  # وارد کردن مدل شبکه عصبی چند لایه  /  Multi-Layer Perceptron (MLP) Modell importieren

# ایجاد مدل شبکه عصبی
mlp_model = MLPRegressor(hidden_layer_sizes=(50,), max_iter=1000, random_state=42)  # تنظیمات مدل  /  Modell konfigurieren

# آموزش مدل روی داده‌های آموزشی
mlp_model.fit(X_train, y_train)  # آموزش مدل با استفاده از داده‌های آموزشی  /  Modell mit Trainingsdaten trainieren

# پیش‌بینی با استفاده از مدل آموزش‌دیده
y_pred_mlp = mlp_model.predict(X_test)  # پیش‌بینی با استفاده از داده‌های تست  /  Vorhersagen mit Testdaten machen

# ارزیابی مدل
mse_mlp = mean_squared_error(y_test, y_pred_mlp)  # محاسبه خطای مربعات میانگین  /  Mittleren quadratischen Fehler berechnen
r2_mlp = r2_score(y_test, y_pred_mlp)  # محاسبه ضریب تعیین  /  Bestimmtheitsmaß berechnen

print(f"Mean Squared Error (Neural Network): {mse_mlp}")  # نمایش خطای مربعات میانگین  /  Mittleren quadratischen Fehler anzeigen
print(f"R² Score (Neural Network): {r2_mlp}")  # نمایش ضریب تعیین  /  Bestimmtheitsmaß anzeigen
```

```
Mean Squared Error (Neural Network): 0.36774858168102
```

[25]:

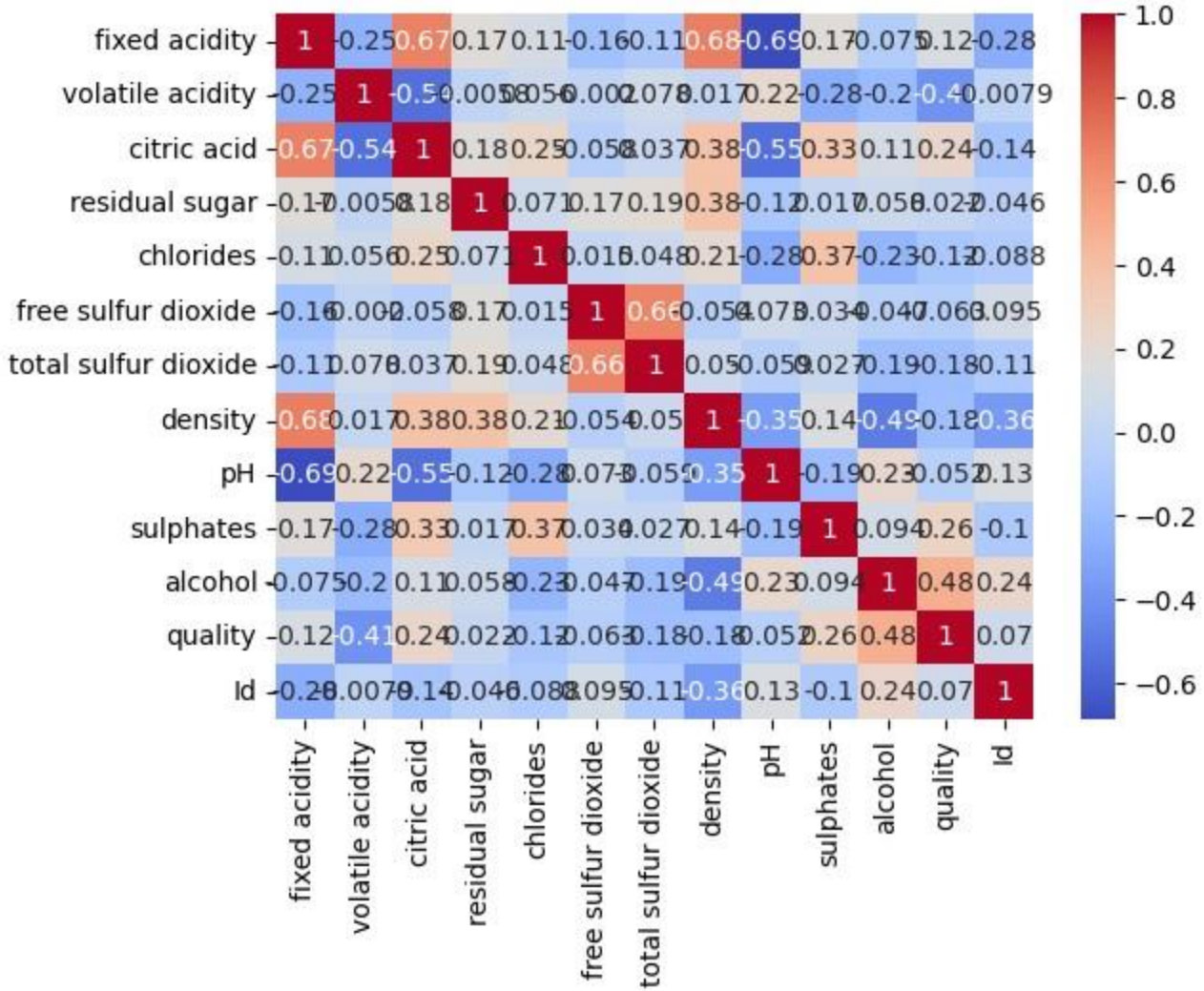
```
# 1. رگرسیون خطی (Linear Regression)
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# ایجاد مدل رگرسیون خطی
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# پیش‌بینی با مدل رگرسیون خطی
y_pred_lr = lr_model.predict(X_test)

# محاسبه R² و میانگین مربعات خطا برای رگرسیون خطی
mse_lr = mean_squared_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)

# 2. جنگل تصادفی (Random Forest)
from sklearn.ensemble import RandomForestRegressor

# ایجاد مدل جنگل تصادفی
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# پیش‌بینی با مدل جنگل تصادفی
y_pred_rf = rf_model.predict(X_test)

# محاسبه R² و میانگین مربعات خطا برای جنگل تصادفی
mse_rf = mean_squared_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

# 3. XGBoost
import xgboost as xgb

# ایجاد مدل XGBoost
xgb_model = xgb.XGBRegressor(n_estimators=100, random_segment=42)
xgb_model.fit(X_train, y_train)

# پیش‌بینی با مدل XGBoost
y_pred_xgb = xgb_model.predict(X_test)

# محاسبه R² و میانگین مربعات خطا برای XGBoost
mse_xgb = mean_squared_error(y_test, y_pred_xgb)
r2_xgb = r2_score(y_test, y_pred_xgb)

# 4. نمایش نتایج
print("Mean Squared Error (Linear Regression):", mse_lr)
print("R² Score (Linear Regression):", r2_lr)
print("Mean Squared Error (Random Forest):", mse_rf)
print("R² Score (Random Forest):", r2_rf)
print("Mean Squared Error (XGBoost):", mse_xgb)
print("R² Score (XGBoost):", r2_xgb)
```

```
Mean Squared Error (Linear Regression): 0.38003245026277527
R² Score (Linear Regression): 0.3170693672733127
Mean Squared Error (Random Forest): 0.2989554585152839
R² Score (Random Forest): 0.462767349736139
Mean Squared Error (XGBoost): 0.36556705832481384
R² Score (XGBoost): 0.3430641293525696
```

5. Ergebnisse und Erkenntnisse

- Ergebnisse:
- Das Random Forest-Modell hat die besten Vorhersagen erzielt.
- Wichtige Merkmale, die die Weinqualität beeinflussen: Alkoholgehalt, Säuregehalt, pH-Wert.
- Schlussfolgerungen:
- Das Random Forest-Modell liefert robuste Vorhersagen und ist für die Weinqualitätsvorhersage am geeignetsten.

[4]:
```
# مرحله 4: تجزیه و تحلیل داده‌ها
# Schritt 4: Daten analysieren
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')  # نقشه حرارتی برای بررسی همبستگی‌ها
# Heatmap zur Überprüfung der Korrelationen
plt.show()
```



6. Visualisierungen

- Power BI:
- Interaktive Dashboards zur Analyse und Visualisierung von Zusammenhängen zwischen den Vari
- Verwendung von Balkendiagrammen, Streudiagrammen und Matrix Visualizations.
- Excel:
- Erstellung von Pivot-Tabellen zur Untersuchung von Datenzusammenhängen.
- Visualisierung durch Balken- und Streudiagramme.
- Python:
- Matplotlib & Seaborn für die Visualisierung der Datenverteilung und der Korrelationen.
- Erstellung von Histogrammen und Korrelationsmatrizen.

Eine Analyse zur Weinqualität • Last saved: Today at 2:18 PM

Search

Sign in

File Home Insert Modeling View Optimize Help Format Data / Drill

Share

Cut Copy Format painter — Clipboard

Get data Excel workbook OneLake catalog SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

New visual Text box More visuals — Insert

New visual calculation New measure Quick measure — Calculations

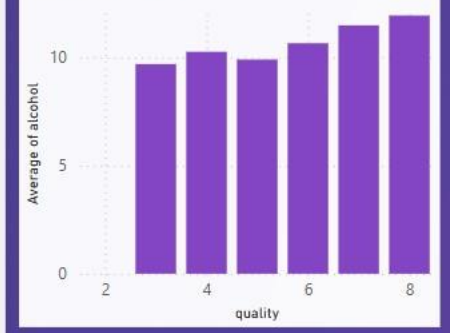
Sensitivity — Sensitivity

Publish — Share

Copilot — Copilot

Eine Analyse zur Weinqualität

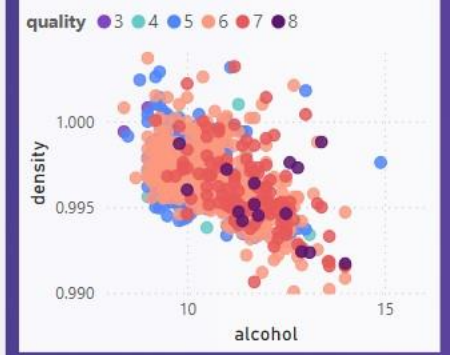
Durchschnittlicher Alkoholgehalt je Qualitätsstufe



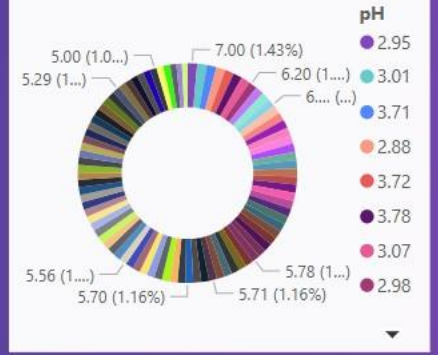
quality: 3 — 8

alcohol: 8.40 — 14.90

quality, alcohol and density



Average of quality by pH



Drill on Columns

alcohol	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0
8.40													
8.50													
8.70													
8.80													
9.00	6.00			4.00					5.00	5.50			
9.10				4.00		5.00							
9.20	5.00		4.00	5.00					5.00		5.00		5
9.23													
9.25													
9.30	5.20		5.00					5.00				5.00	
9.40	5.30		6.00		5.50		5.00		5.00			5.00	5
Total	**5.44**	**5.77**	**5.63**	**5.22**	**5.26**	**5.25**	**5.47**	**5.47**	**5.41**	**5.74**	**5.52**	**5.62**	**5**

Filters

Search

Filters on this visual

- alcohol is (All)
- Average of quality is (All)
- citric acid is (All)
- pH is (All)
- sulphates is (All)

Add data fields here

Filters on this page

Add data fields here

Filters on all pages

Add data fields here

Visualizations

Build visual

Rows: alcohol, sulphates

Columns: citric acid, pH

Values: Average of quality

Drill through

Cross-report On

Keep all filters On

Add drill-through fields here

Data

Search

cleaned_WineQT

- alcohol
- chlorides
- citric acid
- density
- fixed acidity
- free sulfur dioxi...
- Id
- pH
- quality
- residual sugar
- sulphates
- total sulfur diox...
- volatile acidity

Page 1

Page 1 of 1

103%

cleaned_WineQT - Copy - Excel

Search

File Home Insert Draw Page Layout Formulas Data Review View Help Power Pivot PivotChart Analyze Design Format

Share

Cut Copy Format Painter Paste

Aptos Narrow (Body) 10

Wrap Text Merge & Center

General

Conditional Formatting Format as Table

Normal Bad Good Neutral Calculation Check Cell

Insert Delete Format

AutoSum Fill Clear Sort & Filter Find & Select

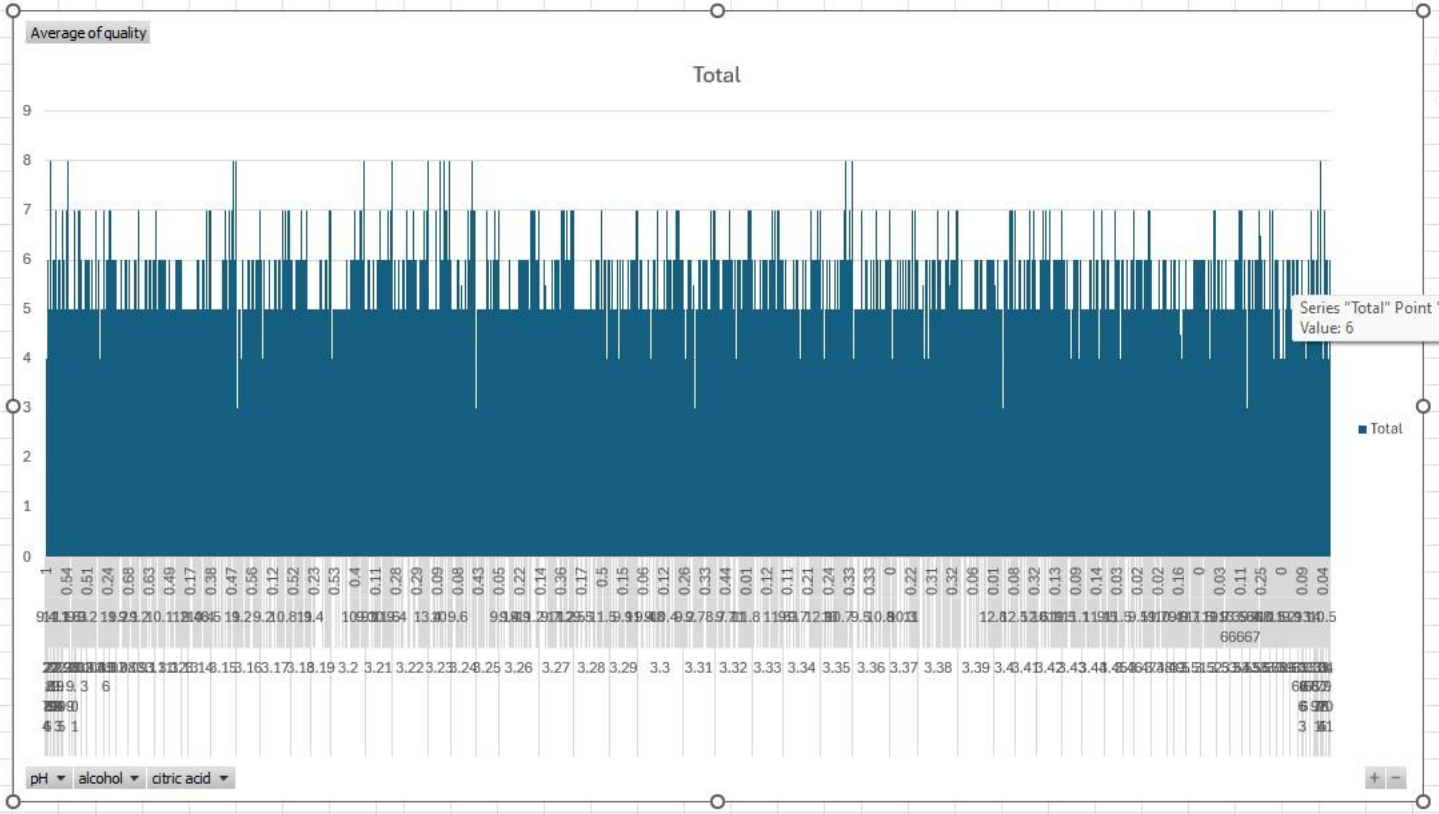
Add-ins

Clipboard Font Alignment Number Styles Cells Editing Add-ins

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again Save As...

Chart 1

Row Labels	Average of quality
2.74	4
9.4	4
1	4
2.86	6
8.4	6
0.74	6
2.88	6.5
9.7	5
0.11	5
9.8	8
0.72	8
2.89	5.333333333
9.6	5
0.39	5
12.7	6
0.37	6
2.9	6
9.1	6
0.68	6
2.92	5.666666667
10.5	7
0.59	7
11.1	5
0.49	5
2.93	6
9.9	6
0.28	6
2.94	5.666666667
9.2	6
0.45	6
9.5	5
0.3	5
2.95	7
11.2	7



PivotChart Fields

Choose fields to add to report:

Search

- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality
- Id

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): pH, alcohol, citric acid

Values: Average of quality

Defer Layout Update Update

Sheet1 cleaned_WineQT - Copy

cleaned_WineQT - Copy - Excel

Search

File Home Insert Draw Page Layout Formulas Data Review View Help Power Pivot Chart Design Format

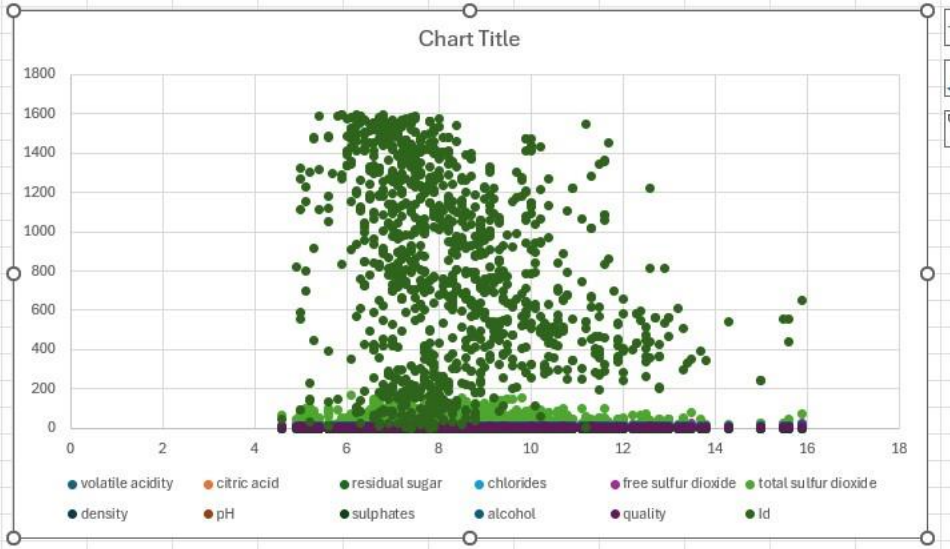
Share

Add Chart Element | Quick Layout | Change Colors | Chart Layouts | Chart Styles | Switch Row/ Column | Select Data | Data | Change Chart Type | Type | Move Chart | Location

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again Save As...

Chart 2

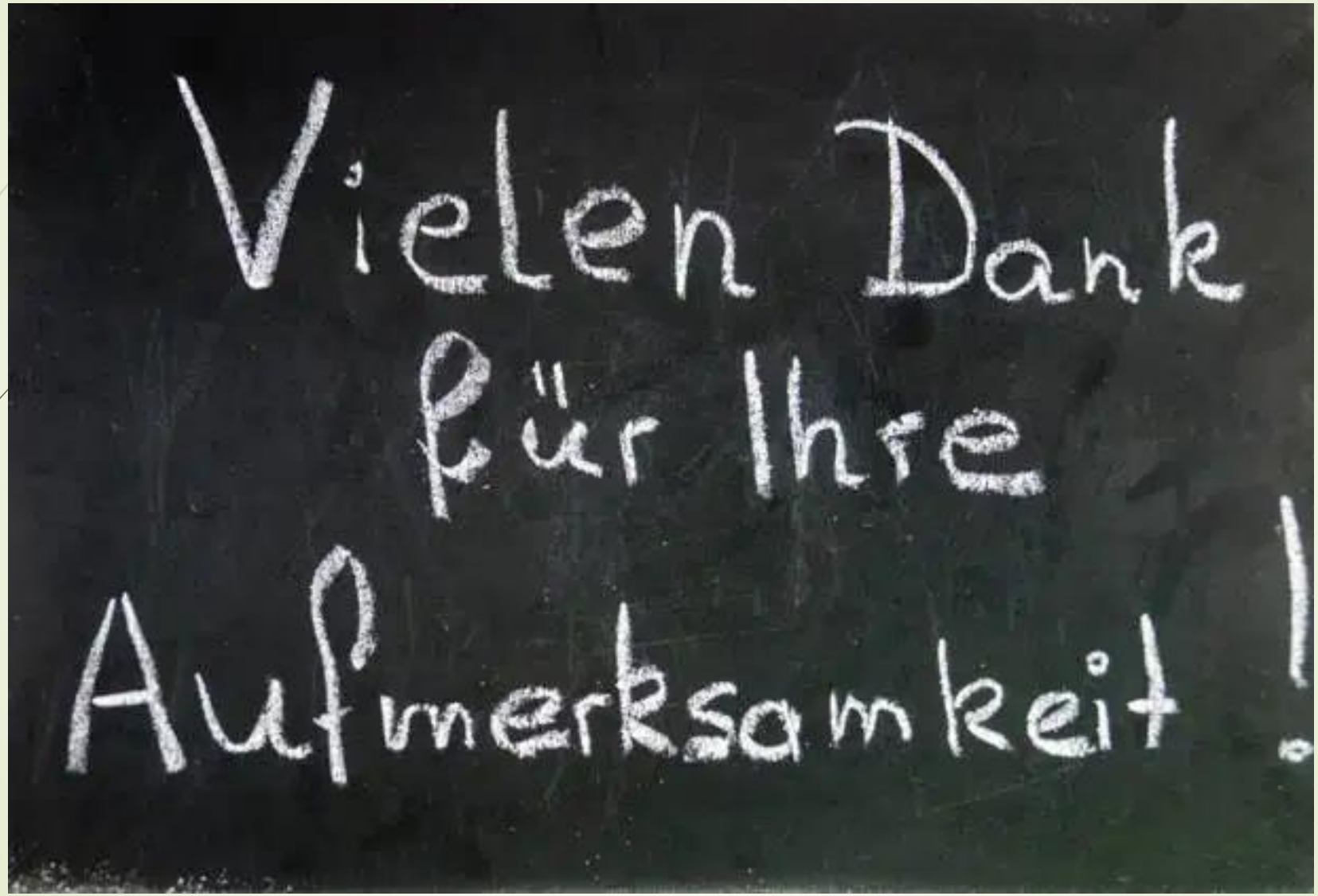
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	fixed acid	volatile ac	citric acid	residual s	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	quality	Id
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	0
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	1
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	2
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	3
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	4
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5	6
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7	8
11	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5	10
12	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5	12
13	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5	13
14	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7	16
15	7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6	19
16	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5	21
17	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5	22
18	8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5	23
19	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6	24
20	6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5	25
21	7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5	26
22	7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5	28
23	7.8	0.645	0	2	0.082	8	16	0.9964	3.38	0.59	9.8	6	29
24	6.7	0.675	0.07	2.4	0.089	17	82	0.9958	3.35	0.54	10.1	5	30
25	8.3	0.655	0.12	2.3	0.083	15	113	0.9966	3.17	0.66	9.8	5	32
26	5.2	0.32	0.25	1.8	0.103	13	50	0.9957	3.38	0.55	9.2	5	34
27	7.8	0.645	0	5.5	0.086	5	18	0.9986	3.4	0.55	9.6	6	35
28	7.8	0.6	0.14	2.4	0.086	3	15	0.9975	3.42	0.6	10.8	6	36
29	8.1	0.38	0.28	2.1	0.066	13	30	0.9968	3.23	0.73	9.7	7	37
30	7.3	0.45	0.36	5.9	0.074	12	87	0.9978	3.33	0.83	10.5	5	40
31	8.8	0.61	0.3	2.8	0.088	17	46	0.9976	3.26	0.51	9.3	4	41
32	7.5	0.49	0.2	2.6	0.332	8	14	0.9968	3.21	0.9	10.5	6	42
33	8.1	0.66	0.22	2.2	0.069	9	23	0.9968	3.3	1.2	10.3	5	43
34	4.6	0.52	0.15	2.1	0.054	8	65	0.9934	3.9	0.56	13.1	4	45
35	7.7	0.935	0.43	2.2	0.114	22	114	0.997	3.25	0.73	9.2	5	46



cleaned_WineQT - Copy

Ready Accessibility: Unavailable

Average: 69.1764668 Count: 14872 Sum: 1027893.12 100%


Vielen Dank
für Ihre
Aufmerksamkeit !