# Term 1 &2 EDA Project

## Wine Quality Dataset

Presented by Clifer Fernandez – April 2020 cohort

# Agenda

- Introduction
- Executive Summary
- Data Profiling
- Data Analysis
- Conclusions
- Recommendations

# Introduction

- Dataset - Properties of Wine
  - 11 physiochemical properties
  - 1 Sensory property : ‘Quality’– based on taste

The Challenge!

- Can we predict the quality of a wine based on these 12 properties?

# Executive Summary

- Original data set – 6497 records
- Analysis indicates
  - No clear combination of properties
  - High quality wines
    - higher alcohol levels
    - lower residual sugar, chlorides and volatile acidity
- Further actions
  - Additional data collection eg: Red / White / Rosé
  - Current analysis and data - select / develop Machine Learning Models that may be able to predict

# Data Profiling

- Data set
  - Import data
  - View records
  - Basic stats
  - Profiling

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| **1** | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| **2** | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| **3** | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |
| **4** | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 | 6497.000 |
| **mean** | 7.215 | 0.340 | 0.319 | 5.443 | 0.056 | 30.525 | 115.745 | 0.995 | 3.219 | 0.531 | 10.492 | 5.818 |
| **std** | 1.296 | 0.165 | 0.145 | 4.758 | 0.035 | 17.749 | 56.522 | 0.003 | 0.161 | 0.149 | 1.193 | 0.873 |
| **min** | 3.800 | 0.080 | 0.000 | 0.600 | 0.009 | 1.000 | 6.000 | 0.987 | 2.720 | 0.220 | 8.000 | 3.000 |
| **25%** | 6.400 | 0.230 | 0.250 | 1.800 | 0.038 | 17.000 | 77.000 | 0.992 | 3.110 | 0.430 | 9.500 | 5.000 |
| **50%** | 7.000 | 0.290 | 0.310 | 3.000 | 0.047 | 29.000 | 118.000 | 0.995 | 3.210 | 0.510 | 10.300 | 6.000 |
| **75%** | 7.700 | 0.400 | 0.390 | 8.100 | 0.065 | 41.000 | 156.000 | 0.997 | 3.320 | 0.600 | 11.300 | 6.000 |
| **max** | 15.900 | 1.580 | 1.660 | 65.800 | 0.611 | 289.000 | 440.000 | 1.039 | 4.010 | 2.000 | 14.900 | 9.000 |

Dataset statistics

| | |
|---|---|
| **Number of variables** | 12 |
| **Number of observations** | 6497 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 1179 |
| **Duplicate rows (%)** | 18.1% |

Variable types

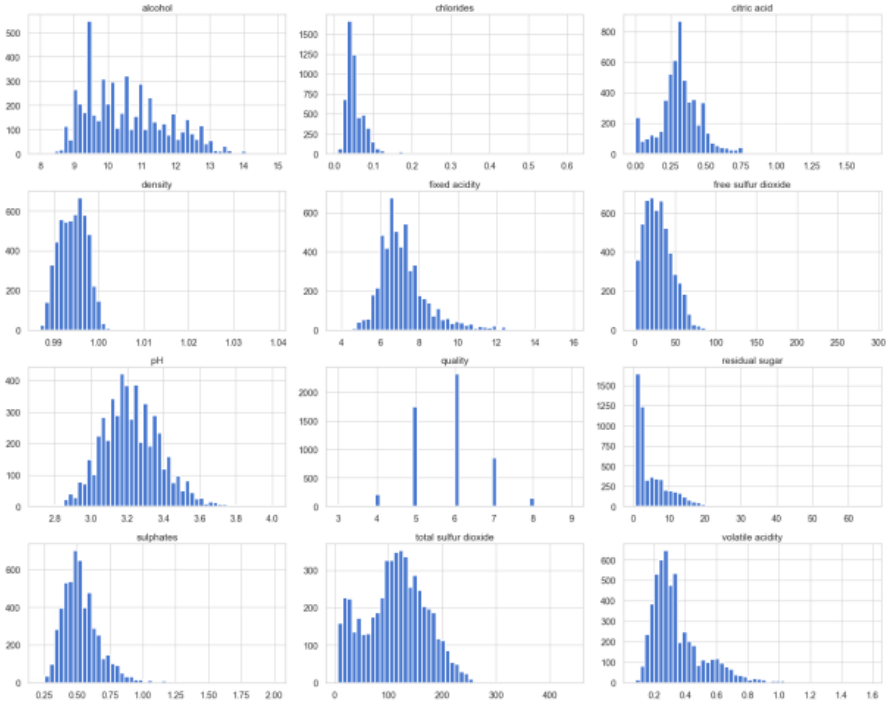
| | |
|---|---|
| **NUM** | 12 |

Dataset has 1179 (18.1%) duplicate rows

`citric acid` has 151 (2.3%) zeros

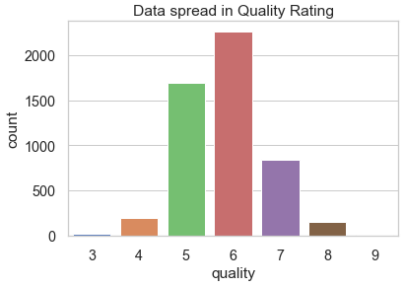
# Data Profiling - Pre-processing

- Pre-processing
  - Remove duplicates
    - Set reduced to 5318 records
  - Check data distribution



# Data Profiling – feature eng.

- Pre-processing
  - Deal with outliers
    - z score  - 4 to +4
    - Removed 155 records
    - Final dataset has 5163 records
- Add feature
  - Wine grade
  - Wine body
    - Link
  - Dryness / sweet
    - Link




| Quality Rating | Count |
|---|---|
| 3 | 20 |
| 4 | 193 |
| 5 | 1690 |
| 6 | 2264 |
| 7 | 845 |
| 8 | 146 |
| 9 | 5 |

| Quality Rating | Group |
|---|---|
| 3,4 | Low |
| 5,6 | Medium |
| 7,8,9 | High |

| Alcohol % | Body |
|---|---|
| <12.5 | Light |
| 12.5 - 13.5 | Medium |
| >13.5 | Full |

| Residual Sugar Content | Dry / Sweet Rating |
|---|---|
| <1 | Bone-dry |
| 1 - 10 | Dry |
| 10 - 35 | Off-Dry |
| 35 - 120 | Sweet |
| 120 - 220 | Very Sweet |

# Data Profiling – Post Profiling

- Post Profiling
  - Rerun stats
    - profiling

Dataset statistics

| | |
|---|---|
| **Number of variables** | 15 |
| **Number of observations** | 5163 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |

Variable types

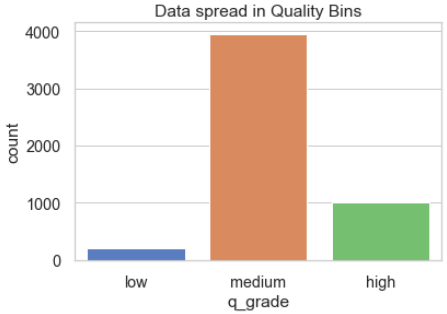
| | |
|---|---|
| **NUM** | 12 |
| **CAT** | 3 |

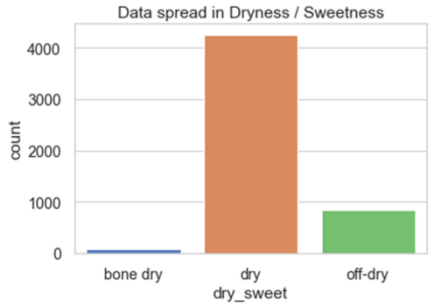
# Data Analysis

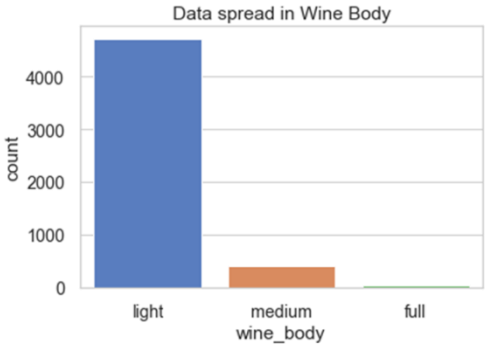
## Data spread – new categories

- Quality grade
  - Low – 213
  - **Medium – 3954**
  - High – 996
- Wine Body
  - **Light – 4721**
  - Medium – 395
  - Full – 47
- Dryness / Sweetness
  - Bone-dry – 72
  - **Dry – 4264**
  - Off-dry – 827

➢ ***Medium grade, Light bodied, Dry wines***



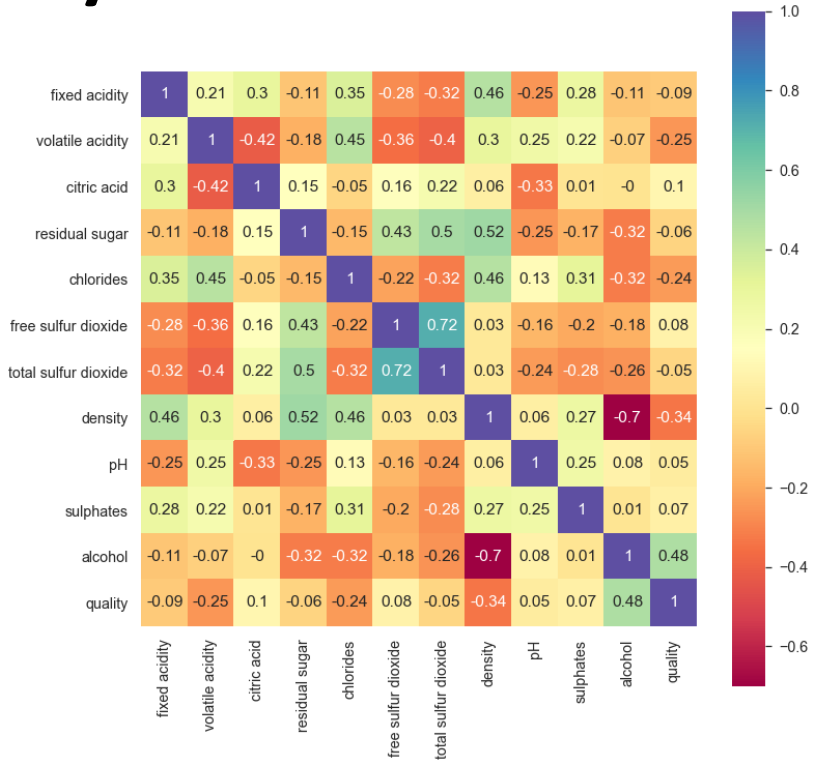



# Data Analysis

**Correlation of the variables**

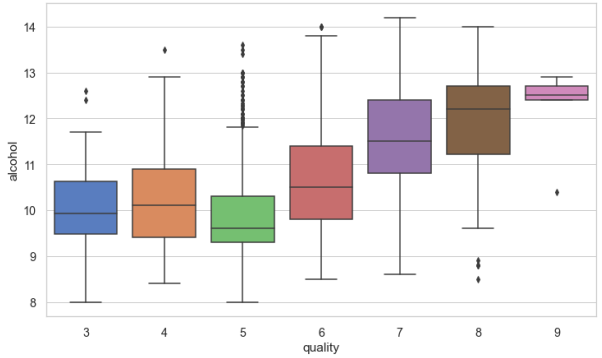
- Quality
  - Alcohol content
  - Density
  - Chlorides (salts)
  - Volatile acidity (acetic acid levels)

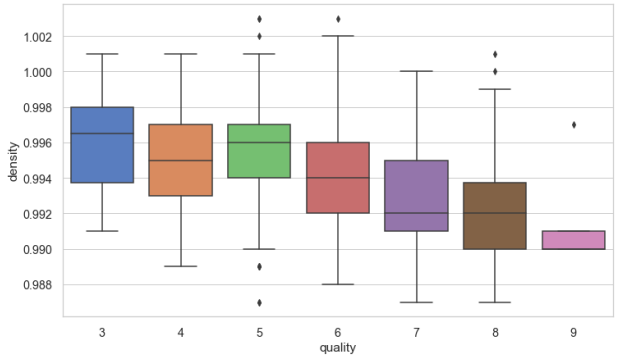


# Data Analysis - Quality

**Variables correlating with Quality**

- vs Alcohol content
  - Higher the quality grade higher the alcohol level
- vs Density
  - Higher the quality grade lower the density

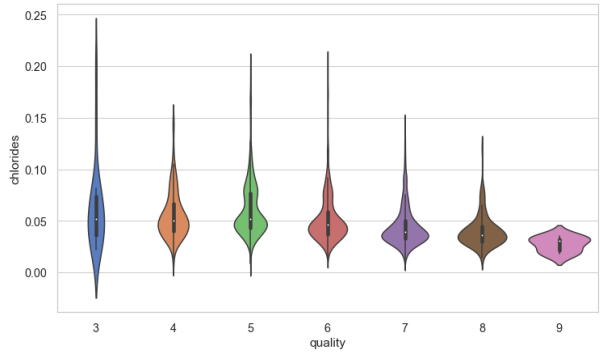


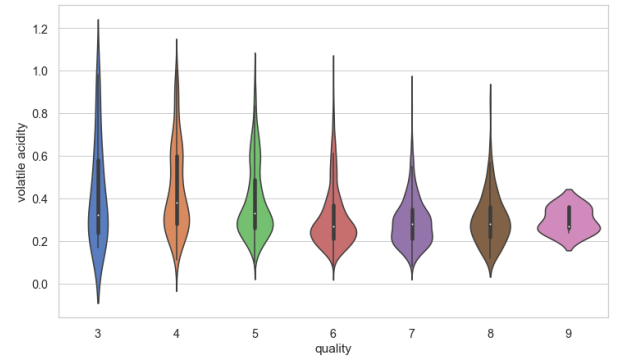



# Data Analysis - Quality

**Variables correlating with Quality**

- vs Chloride levels
  - Higher the quality grade lower the chloride levels
- vs Volatile acidity
  - Higher the quality grade lower the volatile acidity

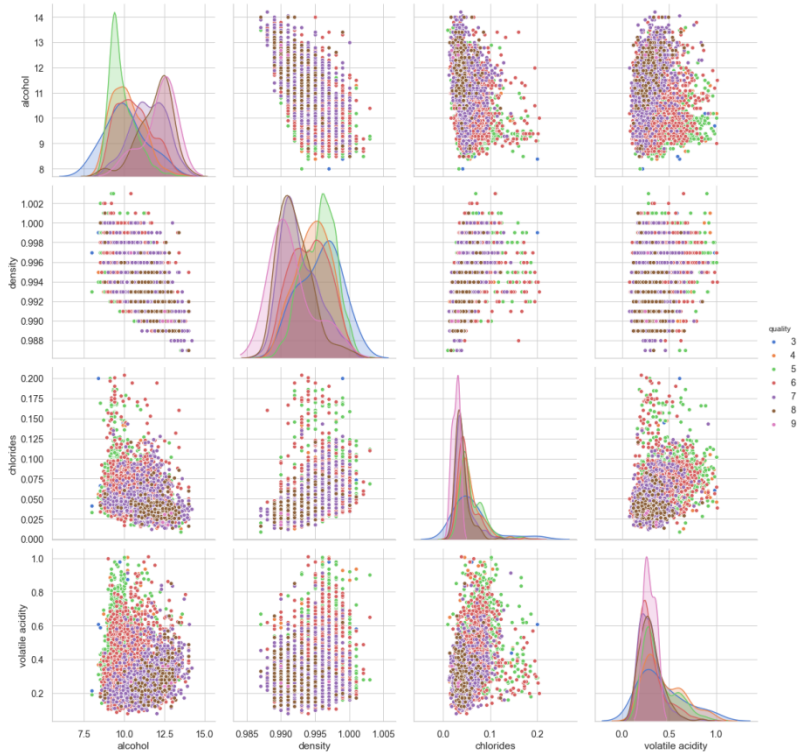


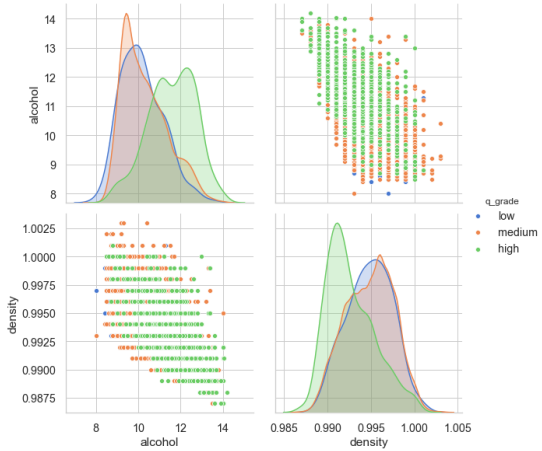
# Data Analysis - Quality

**Variables correlating with Quality**

- Reiterates what we seen
- Note -
  - Chloride and Volatile acidity grouping irrespective of quality grade
    - How much do they really affect the quality rating?

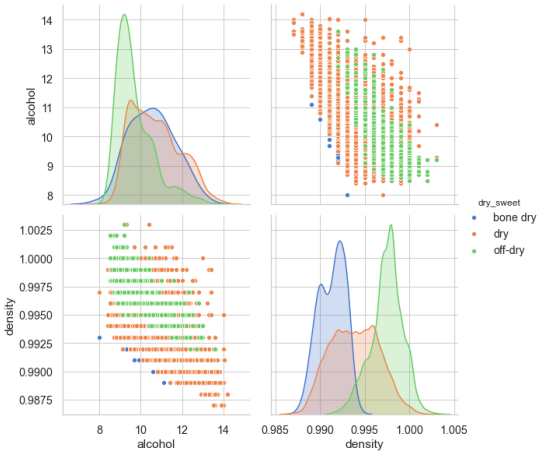


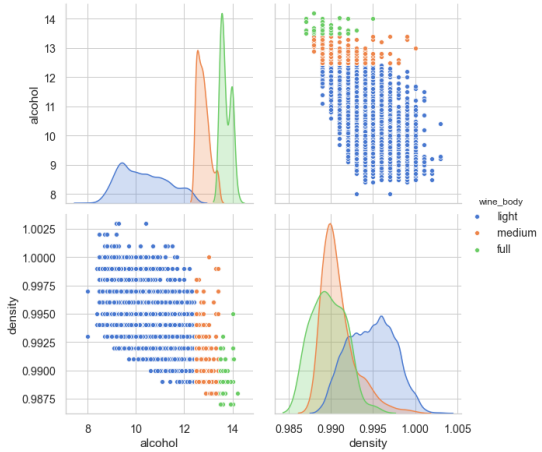
# Data Analysis - Quality




Higher quality wines have higher alcohol content

Sweeter wines have higher density



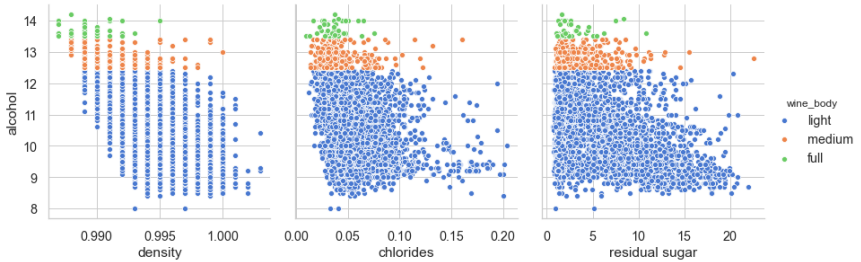




Lighter wines have higher density

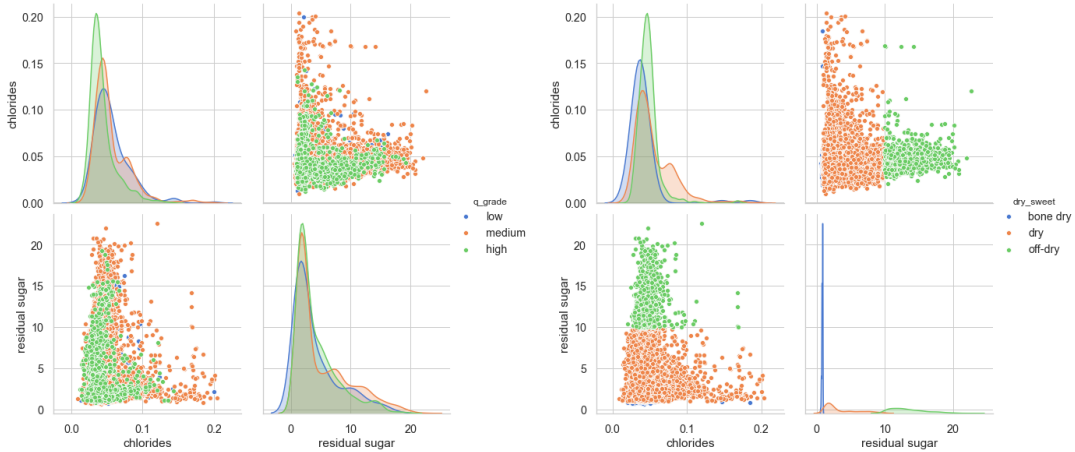
# Data Analysis – Alcohol content

**Variables correlating with Alcohol content**

- vs Density
  - Higher the alcohol content lower the density
- vs Chlorides
  - Low correlation, data is spread out
  - Generally higher the alcohol content lower the chlorides
- Vs Residual sugar
  - Higher the alcohol content lower the residual sugar



# Data Analysis – Alcohol content


The cross plot of residual sugar Vs chlorides

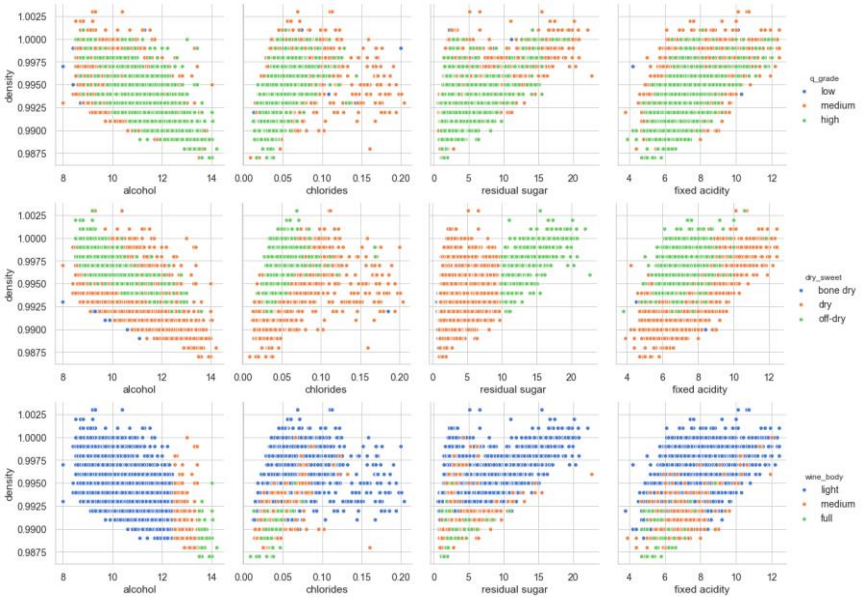
- An obvious 'V' shaped spread
- Could indicate 2 distinct families of wines

Higher residual sugars and higher density are seen in light wines ie: lower alcohol wines

# Data Analysis – Density

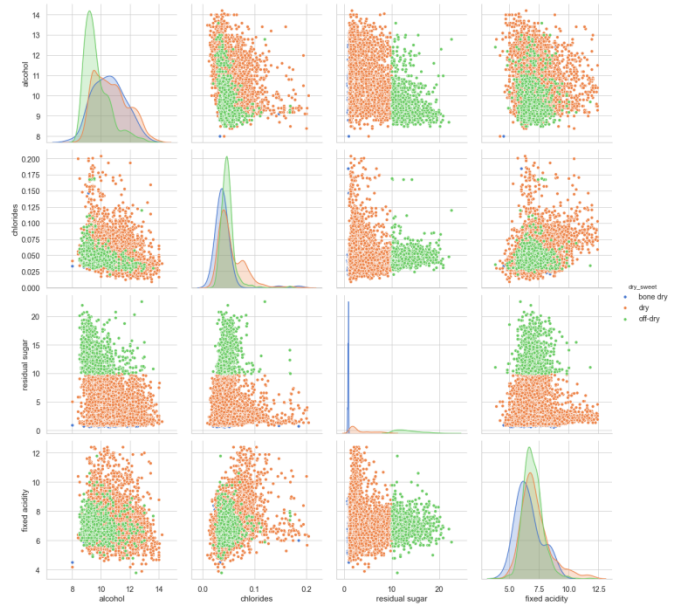
**Variables correlating with Density**

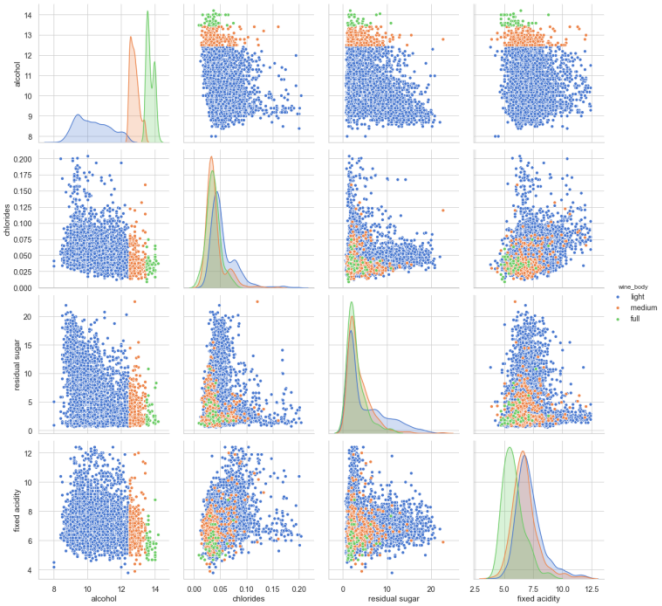
- vs Fixed acidity
  - Higher the density higher the fixed acidity



# Data Analysis – Density

**Variables correlating with Density**

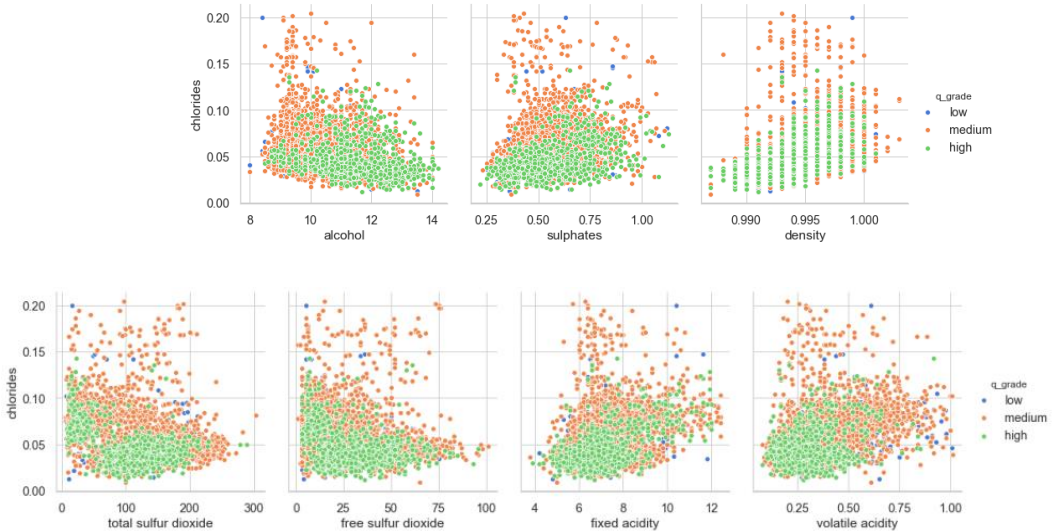


# Data Analysis – Chlorides

**Variables correlating with Chlorides**

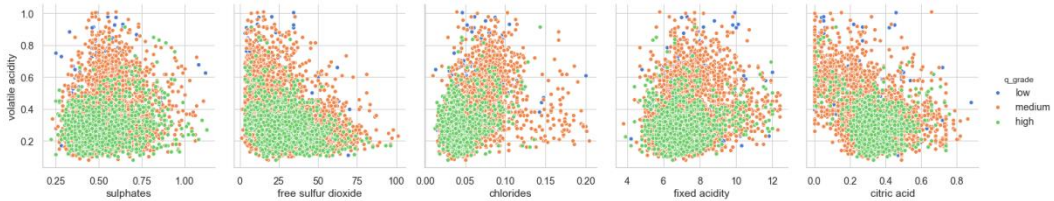
- Data is quiet spread out with weak relationships
- Generally higher grade wines have lower chloride content



# Data Analysis – Volatile Acidity

**Variables correlating with Volatile Acidity**

- Higher quality wines tend to have lower levels of volatile acidity
- Also tend to have lower levels of the ingredients that correlate with Volatile Acidity



# Conclusions

- Dataset concentrated on Medium graded, light bodied , dry wines
- Broadly speaking higher the quality of wine
  - Higher : alcohol content - (human factor?)
  - Lower : density , residual sugar, chlorides and volatile acidity

# Conclusions

- No clear combination of physicochemical properties can be summarised by EDA

# Recommendations

- Data collection
  - Wine type ie: Red, White or Rosé
  - Variety of grape
  - Origin of wine ie: local, region, country etc.
- Current dataset
  - Select / build Machine Learning models to predict quality rating based on the physicochemical properties
  - Additional data from the above collection can be used to improve selected model or select a more appropriate model as necessary

# Thank You!

Santé!