HDSDAJAN19A

**National College of Ireland**

Andrea Mussap, 18177115

Fabio Mussap, 18184618

James Hedderman, 18190481

Lucas Morato, 18184481

July 10th 2019

DATA VISUALIZATION

Continuous assessment – CA1

Table of Contents

[Introduction 2](#_Toc13160100)

[Scope of this project 2](#_Toc13160101)

[Exploratory Data Analysis 3](#_Toc13160102)

[Problem definition 3](#_Toc13160103)

[Data Preparation 3](#_Toc13160104)

[Implementation of the analysis 3](#_Toc13160105)

[Deployment of the analysis 3](#_Toc13160106)

[About the Dataset 3](#_Toc13160107)

[Interactive Visualization 5](#_Toc13160108)

[Objective 5](#_Toc13160109)

[Theme 6](#_Toc13160110)

[Visualizations 6](#_Toc13160111)

[1. Sales by region of origin 6](#_Toc13160112)

[2. Sales by grape variety 8](#_Toc13160113)

[3. Sales per Month 9](#_Toc13160114)

[4. Sales by Occupation (Top 10) 10](#_Toc13160115)

[5. Sales by Age group 11](#_Toc13160116)

[Dashboard 12](#_Toc13160117)

[Complimentary visualizations 13](#_Toc13160118)

[Outliers 13](#_Toc13160119)

[Sales per month by occupations 15](#_Toc13160120)

[Sales by marital status 16](#_Toc13160121)

[Infographic 17](#_Toc13160122)

[Conclusion 18](#_Toc13160123)

[References 19](#_Toc13160124)

[Appendix 20](#_Toc13160125)

[Dataset 20](#_Toc13160126)

[Metadata 20](#_Toc13160127)

# Introduction

This analysis was conducted as part of a group project CA-1 for the Data Visualization module, Higher Diploma in Science in Data Analytics, National College of Ireland.

The goal of this project is to use a large dataset in order to create deliverables using visualization techniques learned in class and in our research.

According to Myatt (2007), a large range of disciplines, from biology to economics, and medical to social research, use data to understand the information available and validate or refute hypotheses.

# Scope of this project

To create the required deliverables:

* Report the exploration and understanding of the data with visual EDA.
* Interactive data visualization
  + Graphics and explanation embedded in this report.
  + Tableau file WineSales\_CA1\_InteractiveViz.twb included in the project submission folder.
  + Available online at [https://public.tableau.com/profile/lucas.morato#!/vizhome/WineSales-InteractiveVisualization/WineSalesOverviewfor2018](https://public.tableau.com/profile/lucas.morato#!/vizhome/WineSales-InteractiveVisualisation/WineSalesOverviewfor2018)
* Infographic
  + WineSales\_CA1\_Infographic.png file included in the project submission
  + This file is also available for download at <https://github.com/andreamussap/NCI_HDSDAJAN19A_DataViz/blob/master/inforgraphic/>

# Exploratory Data Analysis

Exploratory Data Analysis (EDA) provides a structured approach to data analysis. Regardless the particularities of a dataset and research, any EDA should include at least the following general steps:

## Problem definition

The problem that is expected to be solved, and the definition of all the deliverables of the project.

## Data Preparation

The process of collecting, cleaning, and transforming data prior to any analysis or data mining. This is the most time-consuming step, and accordingly with Ruiz (2017) it normally takes around 80% of the working time of a data analyst.

## Implementation of the analysis

Having defined the problem, expected deliverables, and prepared the data, in this step you start the analysis using mathematical and visual tools and appropriate techniques based on the goals of the project.

## Deployment of the analysis

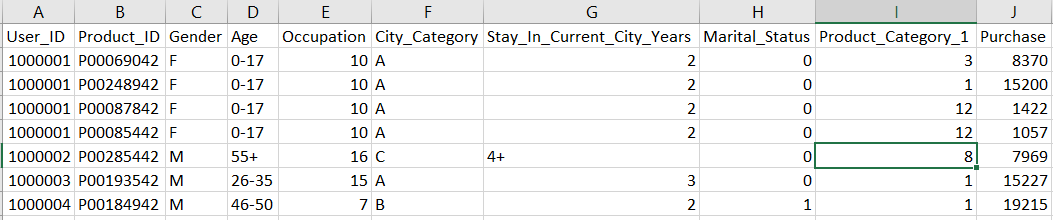
In this last step, the results are made available to the intended audience of the research.

This report follows these four major stages as them provide a logically organized sequence.

# About the Dataset

This project is based on an artificial dataset, originally called Black Friday, created for the educational website Analytics Vidhya, and made available on Kaggle under a CC0: Public Domain license.

After performing some initial EDA on the original dataset, about Black Friday purchases, the group agreed that the original data and datatypes would not generate good visualizations. For example, Occupation, Marital Status, and Product Category variables are represented by numeric values, and City\_Category is represented by only a letter.



Hence, we decided to create a backstory for our project. The story is about wine sales in 2018.

Wine is one of the most popular types of alcohol sold in the world, and France represents more than 10% of world wine consumption and more than 20% of world wine production (Wine Institute, 2019).

(one or two more short paragraphs about wine

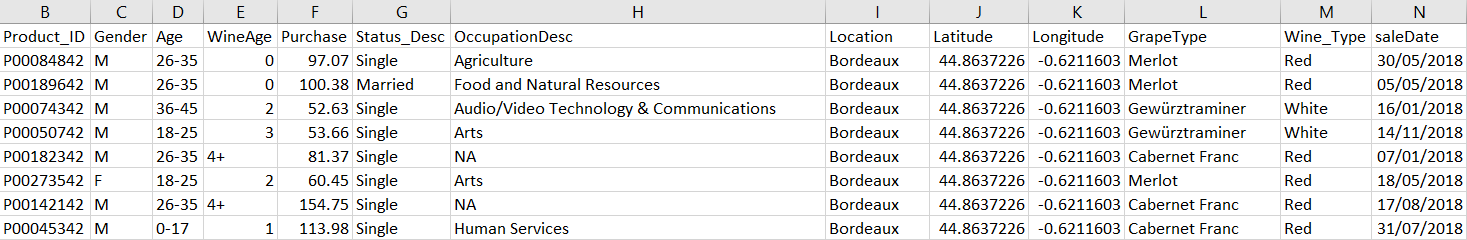
– James can help ??)

* We are a specialist importer of French wine.
* We wish to illustrate our sales data for 2018 – this will be included in our annual report to executive, investors, and shareholders.
* We wish to convey information about the relative popularity of our wines and the breakdown of our customers.

Our wines come from three regions in France, and there are 20 varieties of grapes from what the wines are made of.

This story would make the data more interesting and provide a theme (for example, color palettes, icons, and images) which could be applied across all visualizations.

We modified the dataset then, so that it would represent our sales for 2018 by adding a sales date, mapping category to grape variety, city to wine-region, and so on.



The final dataset, WineSales.csv is included in the project submission, together with the code used to transform the data into the final dataset, file WineSales\_CA1\_DataPreparation.R.

# Interactive Visualization

Once we were satisfied that our data was suitably prepared, we created the initial deliverable, an interactive visualization. This was done in 4 steps:

* Identify the objective and audience.
* Pick a theme (color palette, images).
* Create visualizations (worksheets).
* Create the dashboard and apply filters.

We’ve used Tableau to generate the interactive visualization because it is a powerful tool and it permits to easily publish the dashboard online, as you can see in the following link:

[https://public.tableau.com/profile/lucas.morato#!/vizhome/WineSales-InteractiveVisualization/WineSalesOverviewfor2018](https://public.tableau.com/profile/lucas.morato#!/vizhome/WineSales-InteractiveVisualisation/WineSalesOverviewfor2018)

## Objective

The objective of the visualization is to provide an overview of our company’s sales for 2018.

The visualization has been created with the company directors in mind as a strategic tool which allows them to view:

* The total value of sales for each month.
* The value of sales by region of origin and by the variety of grapes.
* The profile of our customers by age, gender, and occupation.

We believe this provides the audience with valuable information such as:

* Most popular types of wine by gender, occupation, or area.
* Most popular regions based on different customer preferences and profile.
* Value of sales based on preferences of different group ages.

## Theme

After researching other wine related graphical materials, we identified colour schemes commonly associated with wines, such as reds, purples, yellows, browns, and greens. Based on these we created a custom palette comprised of the following:

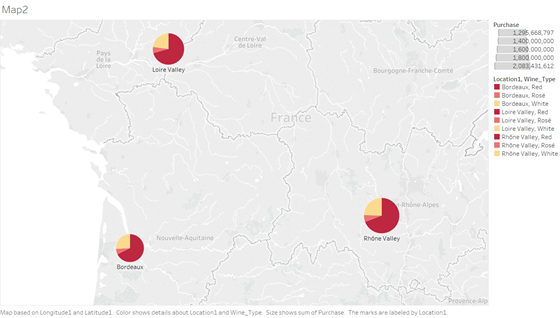
|  |  |  |
| --- | --- | --- |
| Color | Hex Number | Tableau description |
| #290101 color image | #290101 | Very Dark Brown |
| #420101 color image | #420101 | Dark Brown |
| #52151c color image | #52151C | Dark Red |
| #690b22 color image | #690B22 | Dark Red |
| #be2740 color image | #BE2740 | Red |
| #ea7472 color image | #EA7472 | Red |
| #fbda90 color image | #FBDA90 | Light Yellow |
| #caa181 color image | #CAA181 | Light Brown |

## Visualizations

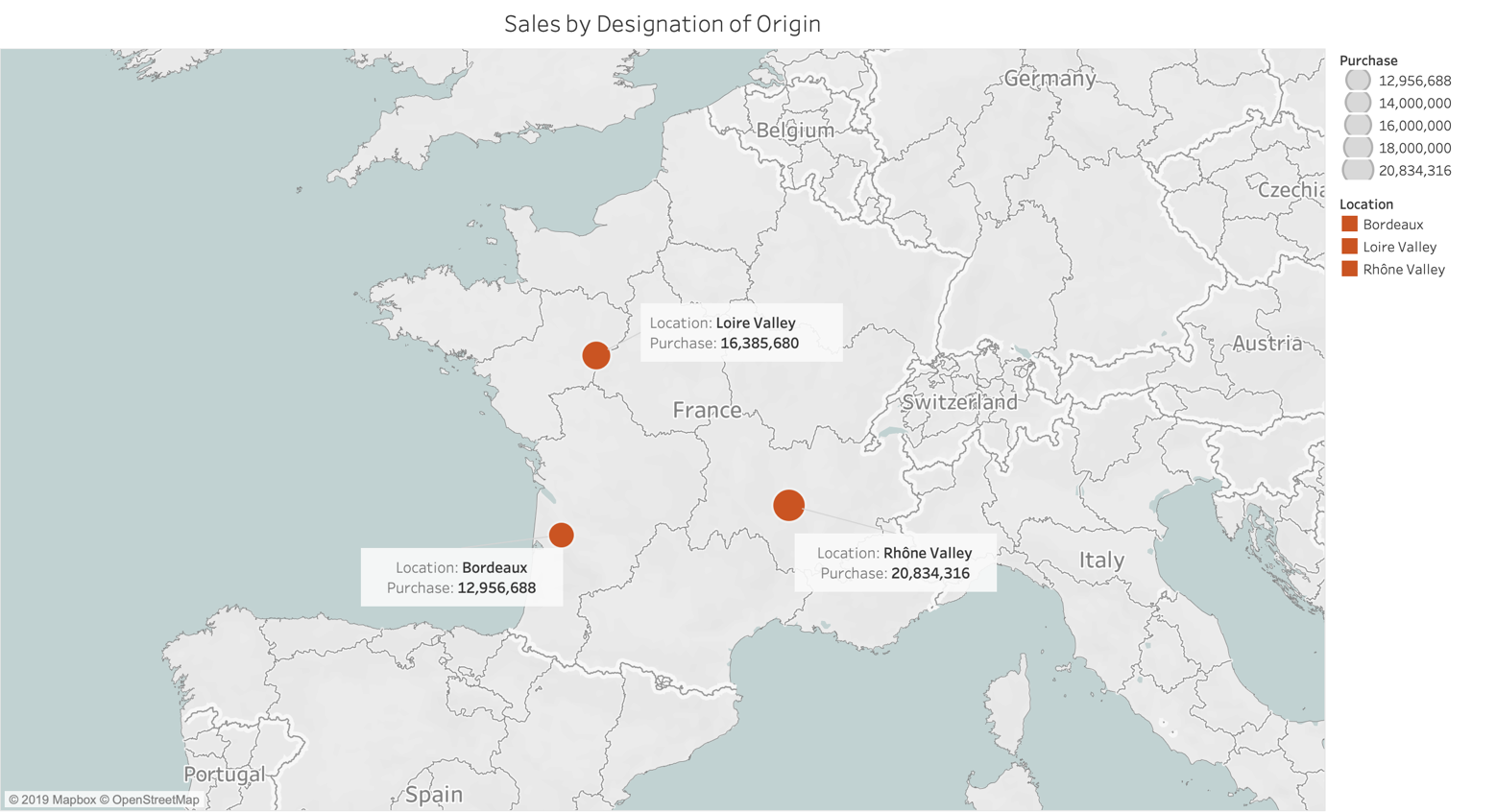
We then created individual visualizations.

### Sales by region of origin

The following map shows the 3 regions from which our products are sourced: Bordeaux, Loire Valley, and Rhône Valley.



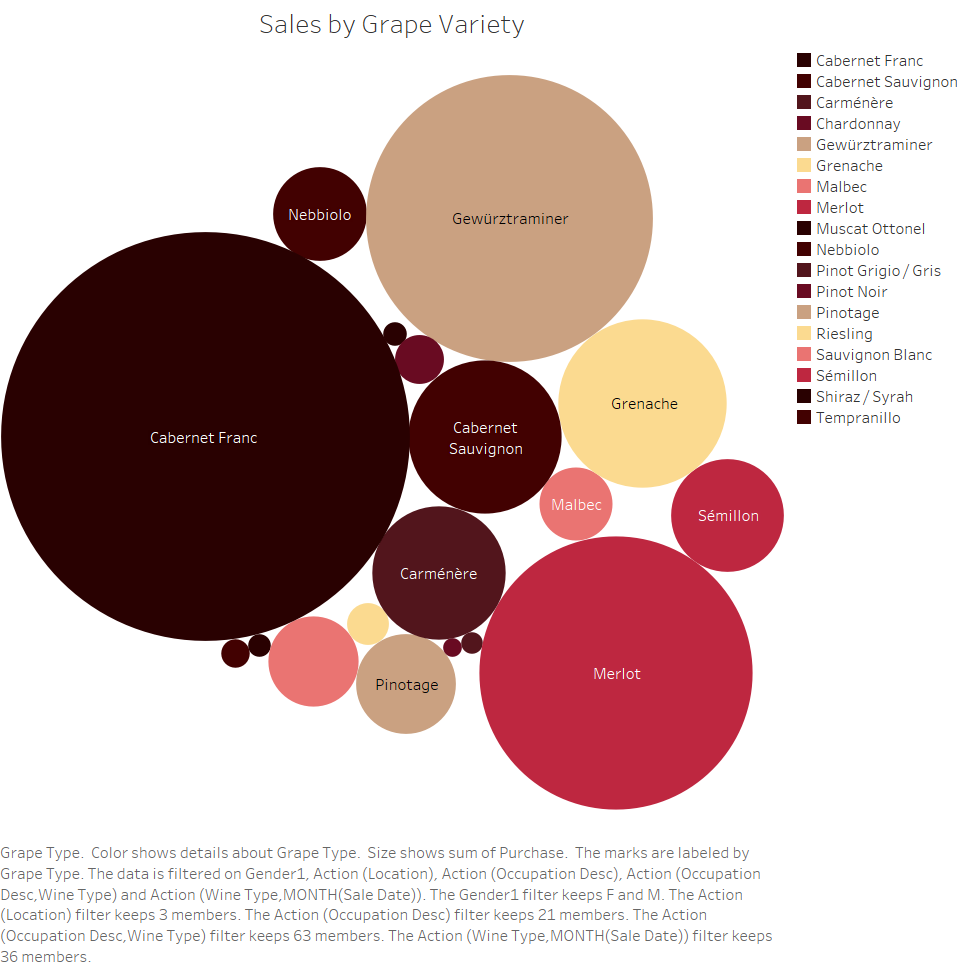
We considered displaying each region as a piechart, however we’ve decided that a simple dot with tooltips for details provided a clearer image.



### Sales by grape variety

We used a bubble plot to show the relative popularity of the grape’s varieties. We chose this format - cluster of bubbles – because it resembles a grapevine.

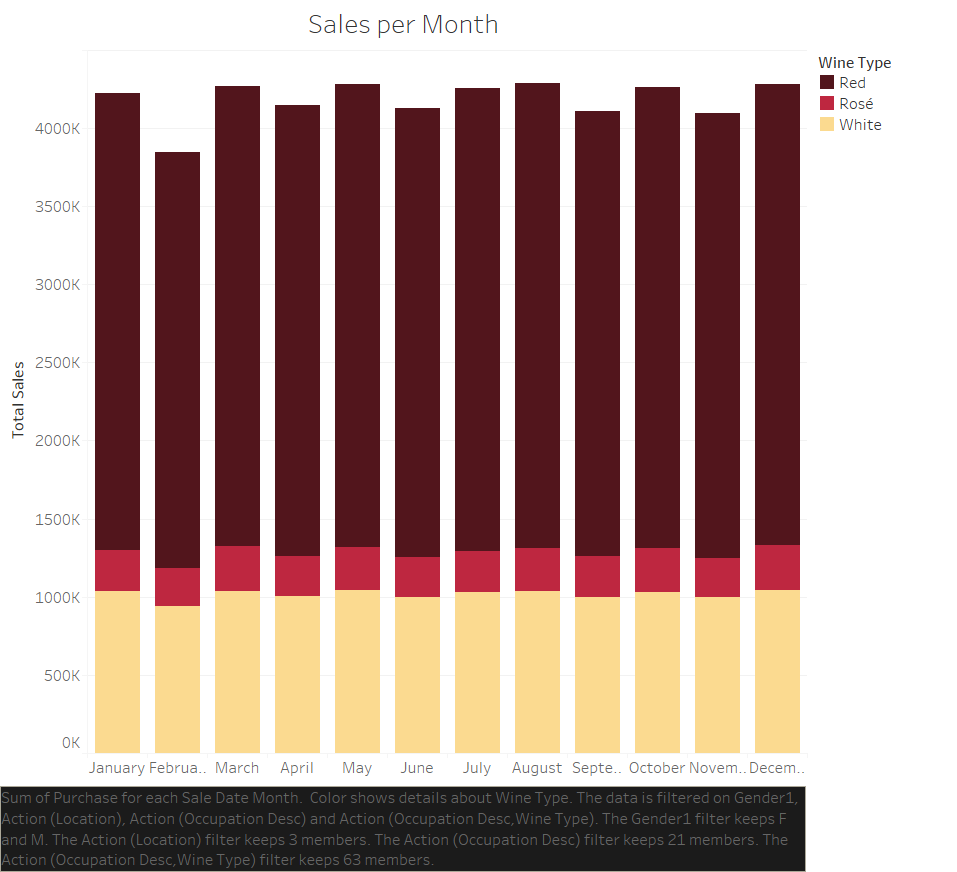
The user can clearly identify the most popular grapes - Cabernet Franc, Merlot, and Gewürztraminer.



### Sales per Month

We’ve used a bar chart to show the sales for each month of 2018 stacked by wine type. We decided for this type of char because it shows a good visualization of the parts of a whole (months of the year), and to track changes over time.

An interesting insight is that we expected to see seasonal variation in the sales of wine types, for example, red during Christmas, rosé in the Summer. However, the Sales-Per-Month chart shows that all types sell consistently across the year, so that assumption was incorrect.



Sum of Purchase for each Sale Date Month. Color shows details about Wine Type. The data is filtered on Gender1, Action (Location), Action (Occupation Desc) and Action (Occupation Desc,Wine Type). The Gender1 filter keeps F and M. The Action (Location) filter keeps 3 members. The Action (Occupation Desc) filter keeps 21 members. The Action (Occupation Desc,Wine Type) filter keeps 63 members.

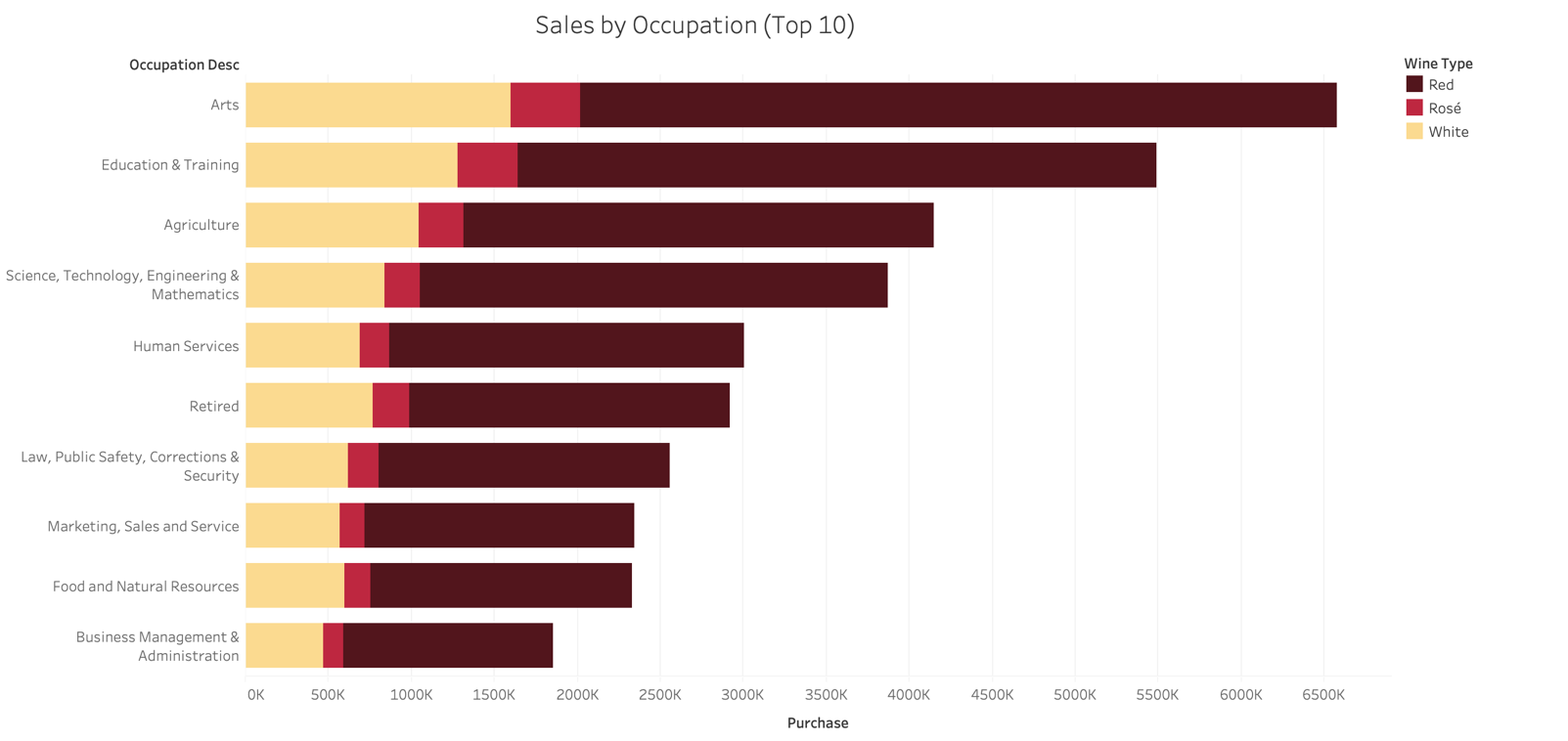
### Sales by Occupation (Top 10)

ACME directors were interested to know more about the type of professionals that purchased our products, as this might be relevant for marketing strategies.

We initially created a heat map to show this information, but we decided then that a horizontal bars chart would show a more interesting view as this type of plot arranges the variables by their lengths, which is proportional to the values that they represent.

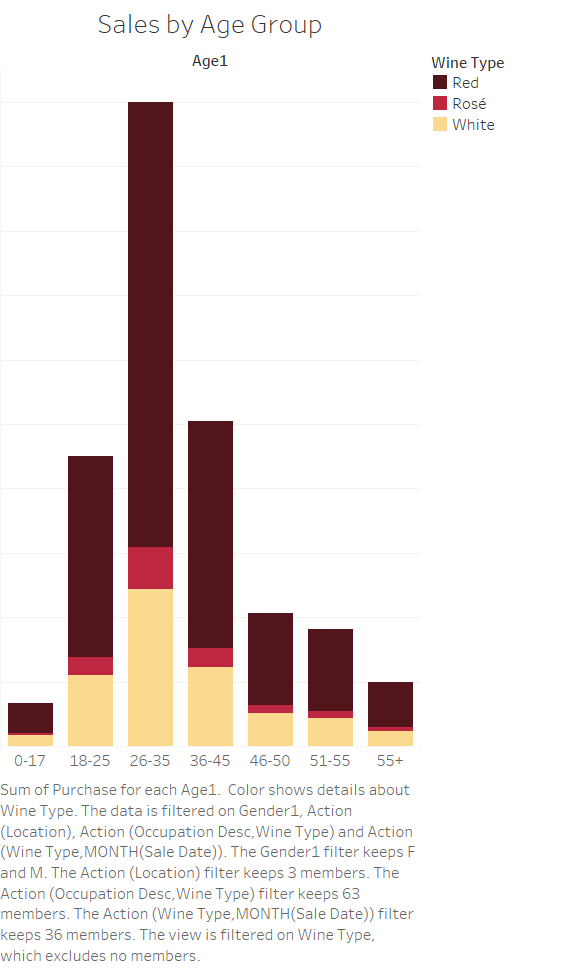
Also, since the list of occupations is long, we’ve agreed that limiting this result to the top 10 occupations would provide more helpful information than showing all of them at once.

This bar chart shows the sales by employment sector.



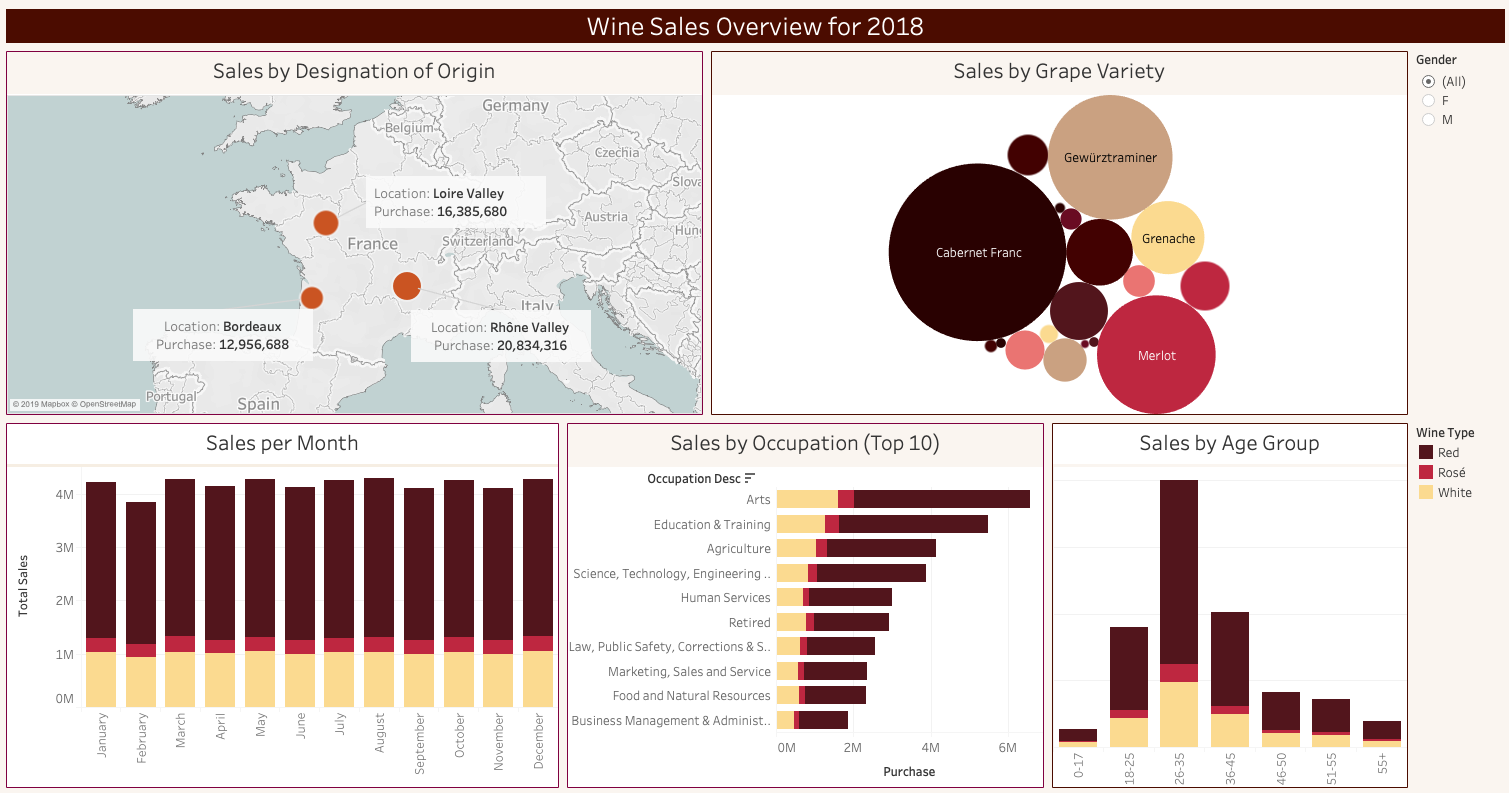
### Sales by Age group

We also charted the sales by age group as a stacked bar chart - our products are surprisingly popular among the 26-35 age range!



# Dashboard

Once the individual sheets were created, we compiled a dashboard in Tableau by arranging on a canvas:



The challenge when creating an interactive dashboard is to connect all the sheets, which makes a consistent theme and narrative essential.

Tableau offers different kinds of interactive filters. The ones we applied in this dashboard were:

* Gender and Wine type boxes: they work independently from each other and can be used within any of the other filters.
* Select filter applied on **Sales by designation of origin** sheet. To apply the filter, select and click on the desired area, then the other 4 sheets will display information regarding just that area.
* Hover filter applied on **Sales per Month** sheet. To see this filter in action, hoover the mouse over the wine types on the stacks, and you’ll see that the rest of the dashboard view will change accordingly with the filter.

The advantage of an interactive visualization is that it allows the user to change the data displayed based on the desired criteria, offering immediate response to different questions.

# Complimentary visualizations

We decided that the interactive visualization should be the main source of information to stakeholders, as the Dashboard can provide the more relevant information in a friendly and meaningful way.

We also designed some complimentary visualizations though, in order to confirm or supplement some of our findings.

## Outliers

From some of the previous visualization we’ve noticed that the distribution of the data for some subsets could be quite spread depending on the filter applied to it.

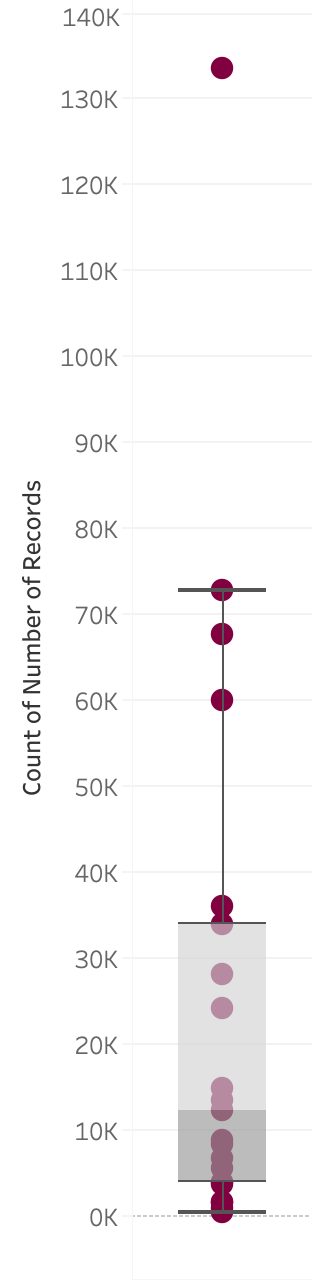
As we’ve seen in the **Sales by Age group** previous visualization, the 26-35 age range’s purchase is way higher than all the other age ranges. So, we’ve decided to use some box-and-whisker plots to show the distribution of the values along the axis, and try to spot some outliers.

The following box-and-whisker plots show the distribution of the number of records for all type of wines by age groups.

Observe in Image\_1 how the total number of records goes beyond than the 4th quartile of the whisker plot as an effect of the 26-35 age group. Then, compare it with Image\_2 and observe that, although we still have an outlier (age group 36-45), its count is not that far from the edge of the whisker plot.

The purpose of this analysis is to verify what we’ve notice already in a previous visualization, and provide more details about the distribution of age-groups subset.

(Image\_1) Count of Number of Records for Wine Type by Age. The Wine Type filter keeps Red, Rosé, and White. The Age filter keeps all of the 7 ranges.



Age: 26-35

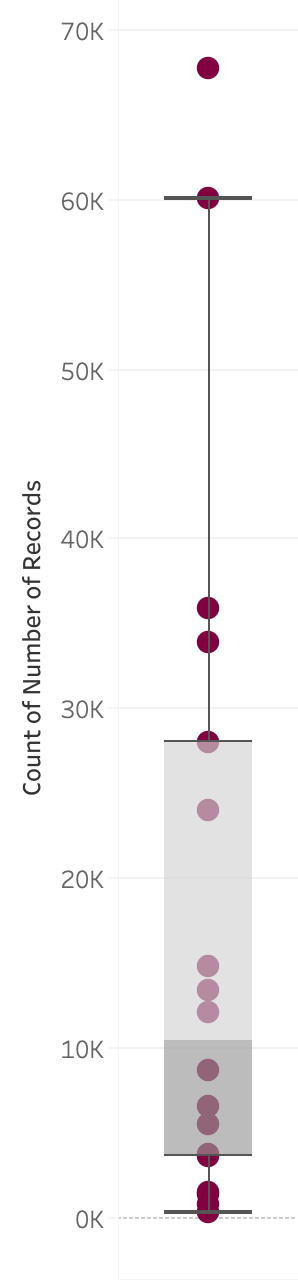
Wine type: Red

Count: 133,562

Age: 36-45

Wine type: Red

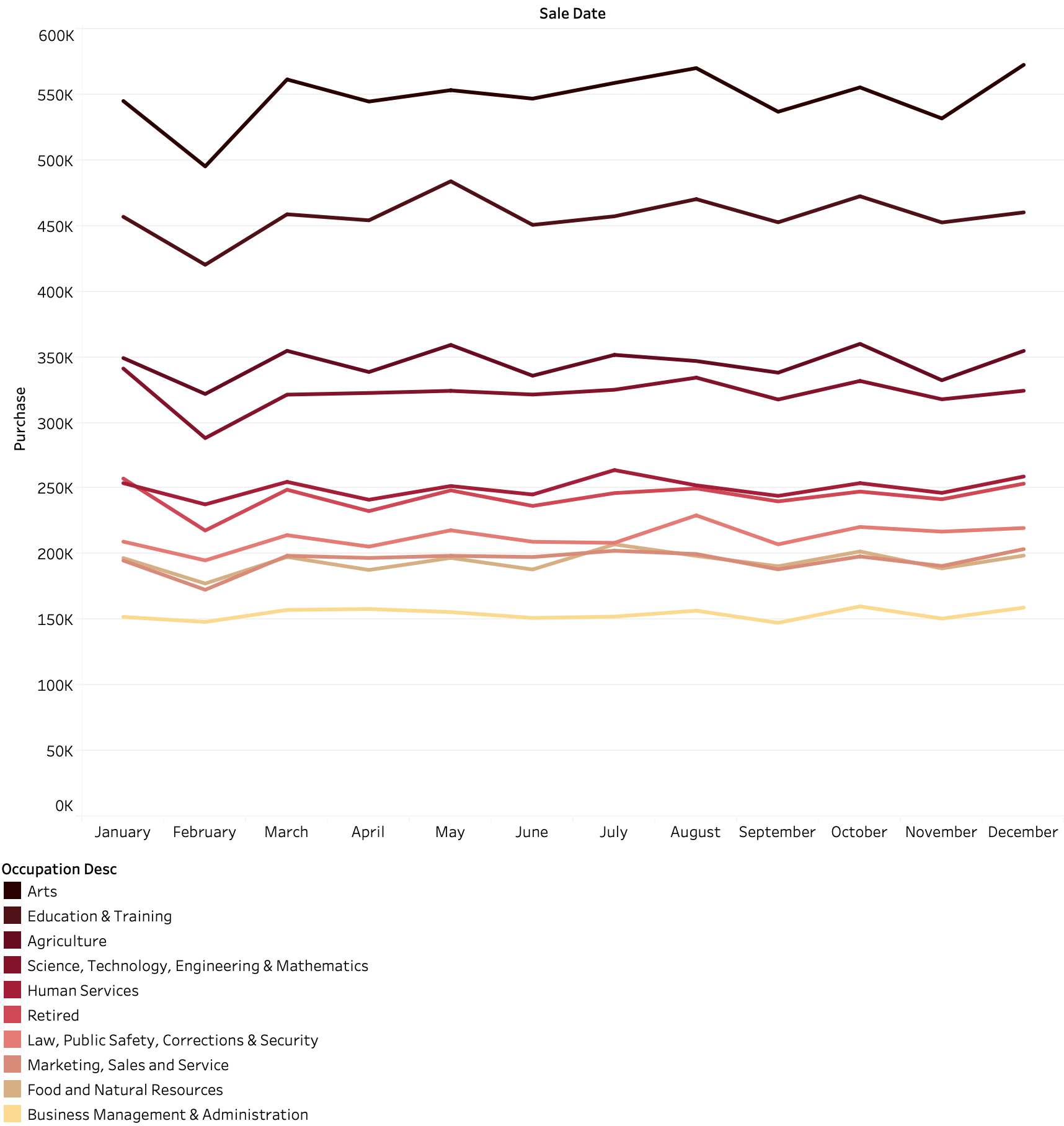
Count: 67,706



(Image\_2) Count of Number of Records for Wine Type by Age. The Wine Type filter keeps Red, Rosé, and White. The Age filter excludes 26-35.

## Sales per month by occupations

Although we’ve presented the total ‘Sales by Occupation (Top 10)’ in the interactive visualization, we decided that it could be helpful to show this same information along the year. For that, we’ve used the lines (time series) type of graphic, which displays the variation of the values of a variable in a timeline.

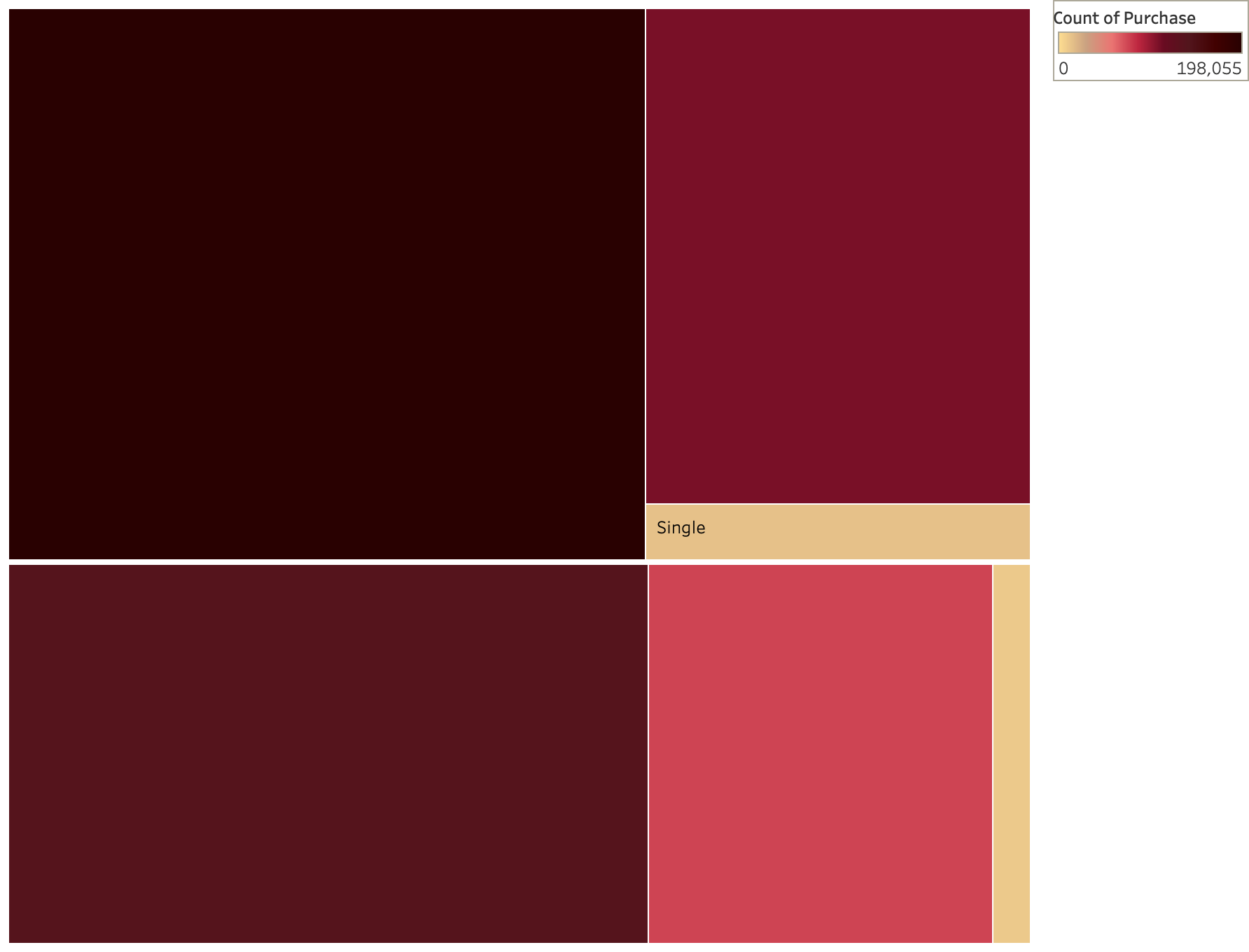
The follow time series graph shows not only the top 10 occupations sales by month, but also their behaviour along the year.

## Sales by marital status

Another interesting insight from this analysis was to notice that single customers buy way more, in all of the three categories of wine, than married customers, particularly for the Red type.

This information is important because it shows to ACME company how their sales is segmented by marital status. Hence, they can think of ways to increase their sales to these two categories of customers.

To show this insight we’ve chosen the Treemap chart because it presents an overview of the structure of the data in glance, by showing a comparison of the proportions between categories accordingly with the size of their area.



Red

Single

36.8%

White

Single

20%

Red

Married

25.6%

White

Married

13.8%

Rosé 2.2%

Rosé Married 1.5%

# Infographic

Infographics are what best represents the proverb "A picture is worth a thousand words". They are static graphics that aim to explain concepts, ideas, or stories graphically. An infographic shows your data by way of pictures, more than by text.

The infographic in this assignment displays a collection of images that tell the story of ACME wine sales in 2018. Its goal is to engage our customers by communicating relevant information in a quick and easy way.

Therefore, the Wine Sales Infographic file (uploaded with this report) informs about ACME sales by gender, age range, top 10 occupations, month-by-month, and so on. It's an overall picture of what we've detailed already along the report.

**Note** - As per Lecturer’s instruction, the WineSales\_CA1\_Infographic.png file is not added to this report, but included in the project submission instead.

# Conclusion

(Aim to create a conclusion/recommendation with this project)

This report is the result of a data analysis visualisation assignment. During its development

# References

Dataset

Wine Institute (2019). ‘World Wine Consumption’. *Wine Institute*. Available at: <https://www.wineinstitute.org/files/World_Consumption_by_Country_2017.pdf> [Accessed 11 June 2019].

Ruiz, A. (2017) ‘The 80/20 data science dilemma’. *Infoworld*,26 September. Available at <https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html> [Accessed 6 June 2019].

Datanovia (2018) *Ggplot colors best tricks you will love.* Available at: <https://www.datanovia.com/en/blog/ggplot-colors-best-tricks-you-will-love/> [Accessed 9 June 2019].

# Appendix

## Dataset

The final dataset, WineSales.csv is included in the project submission, together with the code used to transform the data into the final dataset, file DataPreparation.R.

## Metadata

Description of the variables of the Wine sales dataset. (verify that this is correct !!)

* User\_ID = Unique identifier for each customer.
* Product\_ID = Unique product identifier.
* Gender = M for Male, F for female.
* Age = 5 groups: 0-17, 18-25, 26-35, 36-45, 46-50, 51-55, 55+
* WineAge = How many years the wine has been “aged”.
* Purchase = Value of each purchase in Euro.
* Status\_Desc = Buyer marital status (Single or Married).
* OccupationDesc = Profession of the buyer.
* Location = The region from where the wine was sauced.
* Latitude = Latitude, used for mapping Location.
* Longitude = Longitude, used for mapping Location.
* GrapeType = Variety of grape used in the wines.
* Wine\_Type = Classification of the wine (Red, Rosé, or White).
* saleDate = date of the sale of the wine. Format YYYY/MM/DD.