CS555 PROJECT – FALL 2021

CUSTOMER PERSONALITY ANALYSIS

Rohan J Aditya

U17792248

Project Description

Customer personality analysis is a vital part of businesses. It helps in understanding target audiences in order to increase sales and overall profit margins.

In this project, I analyze whether there is a difference in the amount spent on wine by people grouped by level of education. If the global F test is significant, I also examine which of the pairwise group means are different. I also test whether the population means of the amount spent on wines across people grouped by education level is equal after controlling for income.

Data Set Overview

The data set is a result of five marketing campaigns that surveyed customers and recorded their personal details and the amounts spent by them on various product categories.

The link to the original data set is given below.

https://www.kaggle.com/imakash3011/customer-personality-analysis

Preprocessing and Cleaning

Rows which contain NULL values are removed from the data frame.

Observation of which Education falls into the *Basic* category are also removed because of a lack of data.

Columns which are not relevant to the analysis are also removed.

The cleaned data frame consists of the following columns.

- **Education**: Level of education
- **MntWines**: Amount spent on wines
- **Income**: Yearly household income

Exploratory Data Analysis

1. Number of observations in each group

Group (Education Level)	Number of observations
2nd Cycle	200
Graduation	1116
Master	365
PhD	481

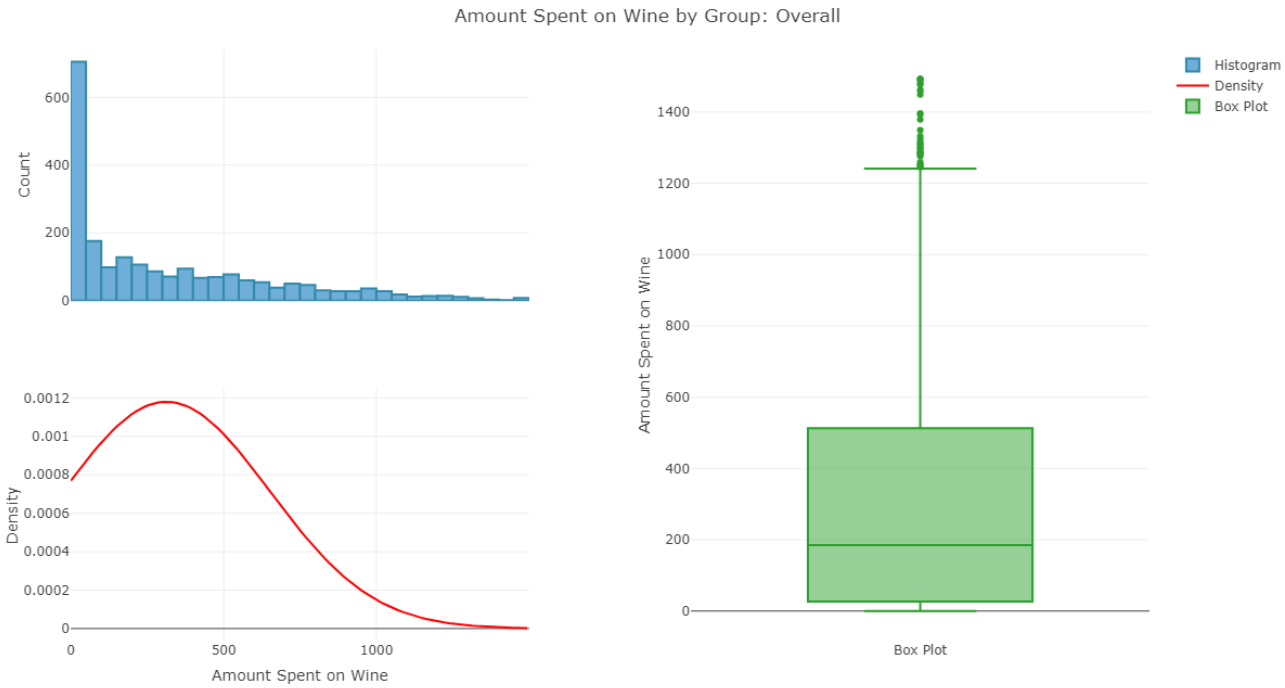
2. Summary Statistics

	2nd Cycle	Graduation	Master	PhD
Minimum	0	0	2	2
First Quartile	9	23	37	52
Median	51	184	177	284
Mean	200.845	285.046	332.981	407.223
Third Quartile	349.25	459.25	544	707
Maximum	1215	1492	1486	1493
Std. Deviation	262.519	308.247	356.120	391.201

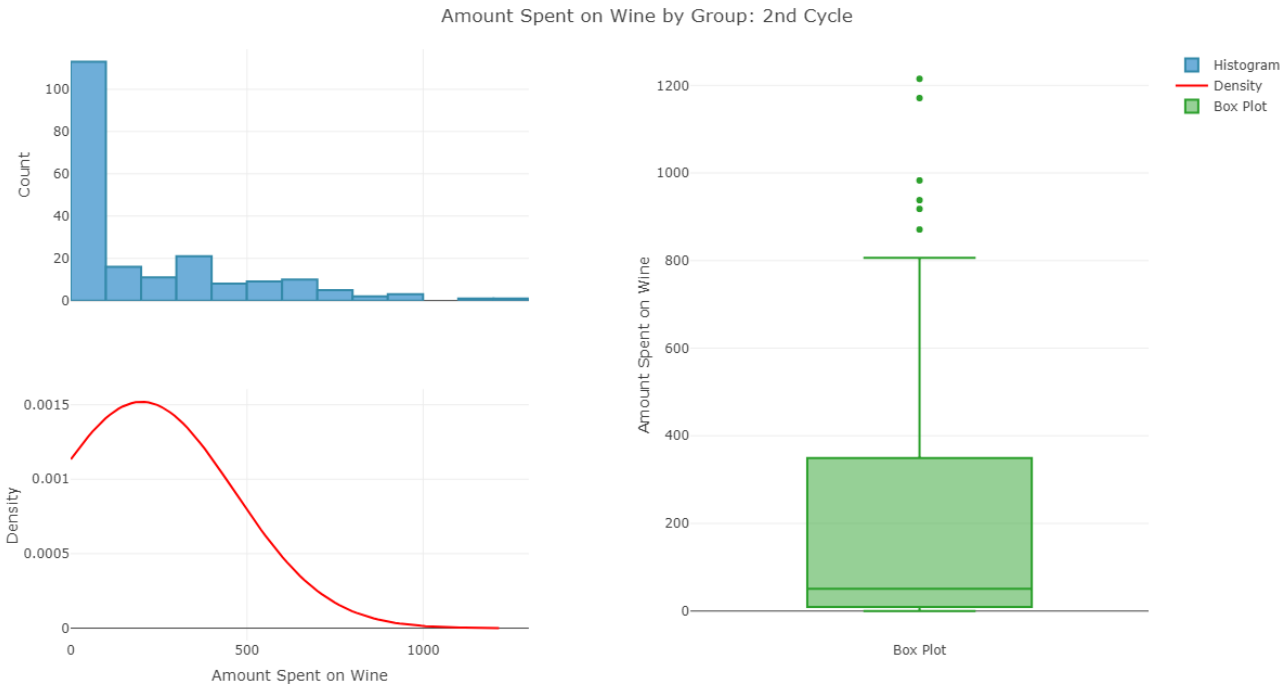
3. Histogram, Density and Box Plots

A histogram, density plot, and box plot of the amount spent on wines by each group and the overall sample is given below.

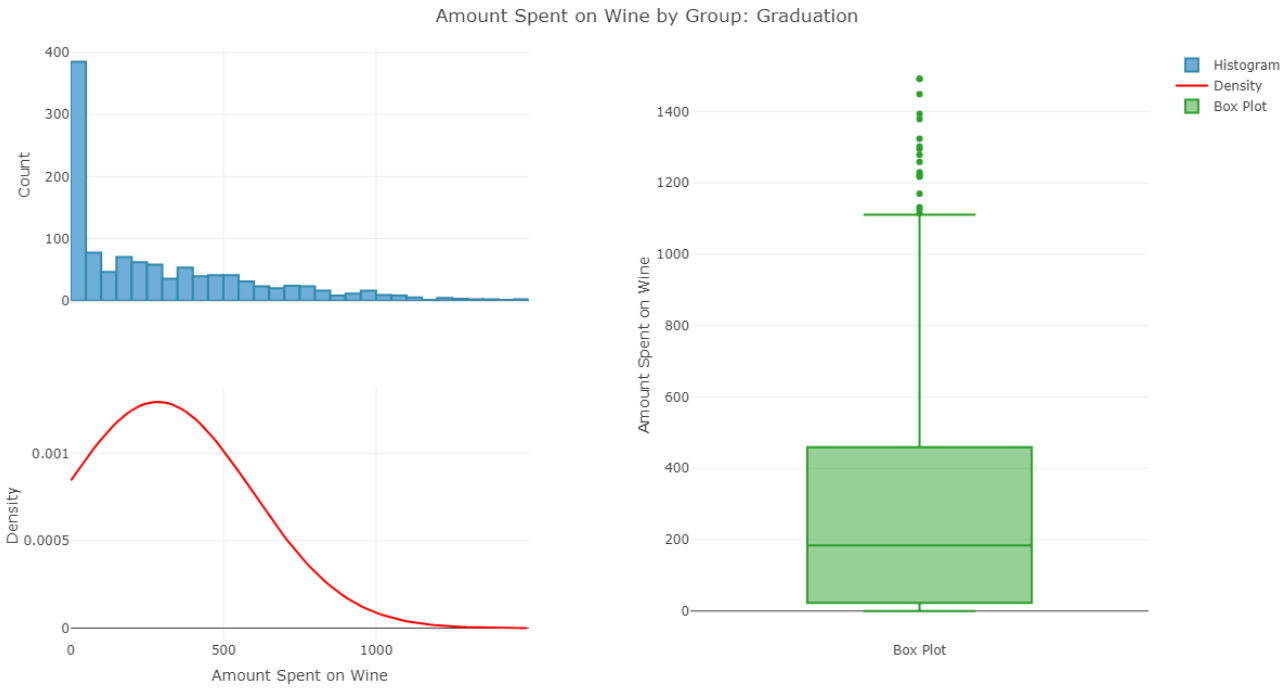
Overall:



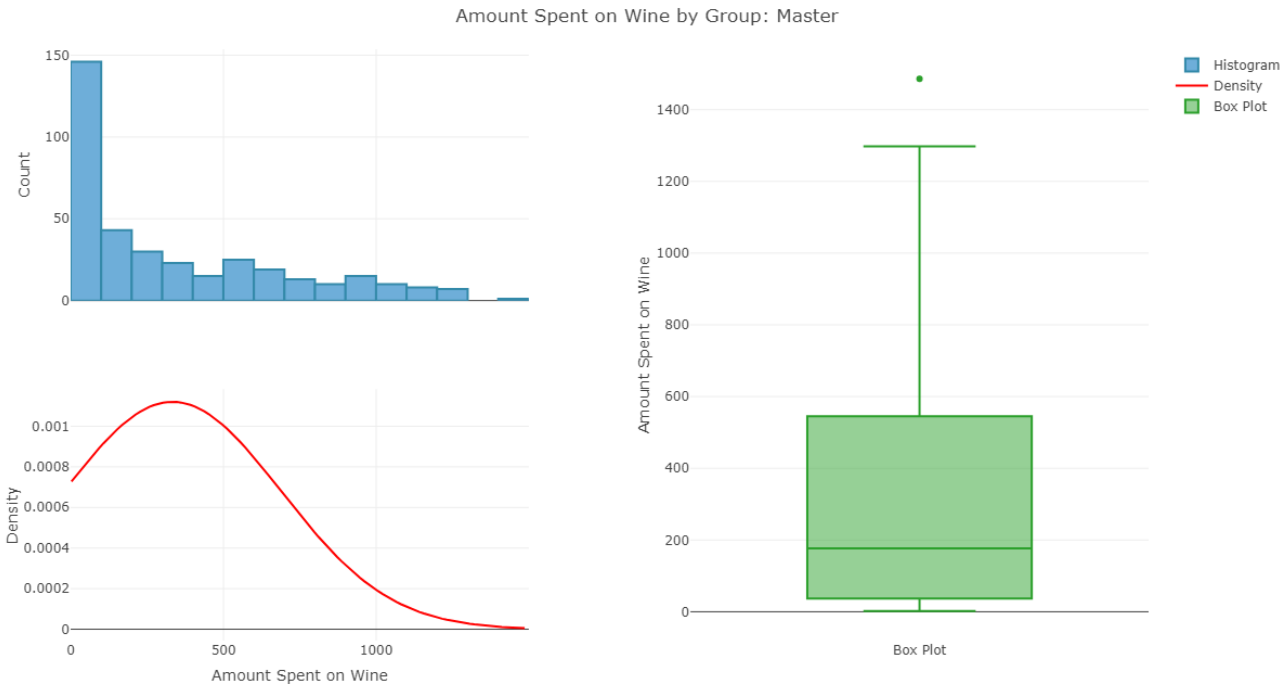
2nd cycle:



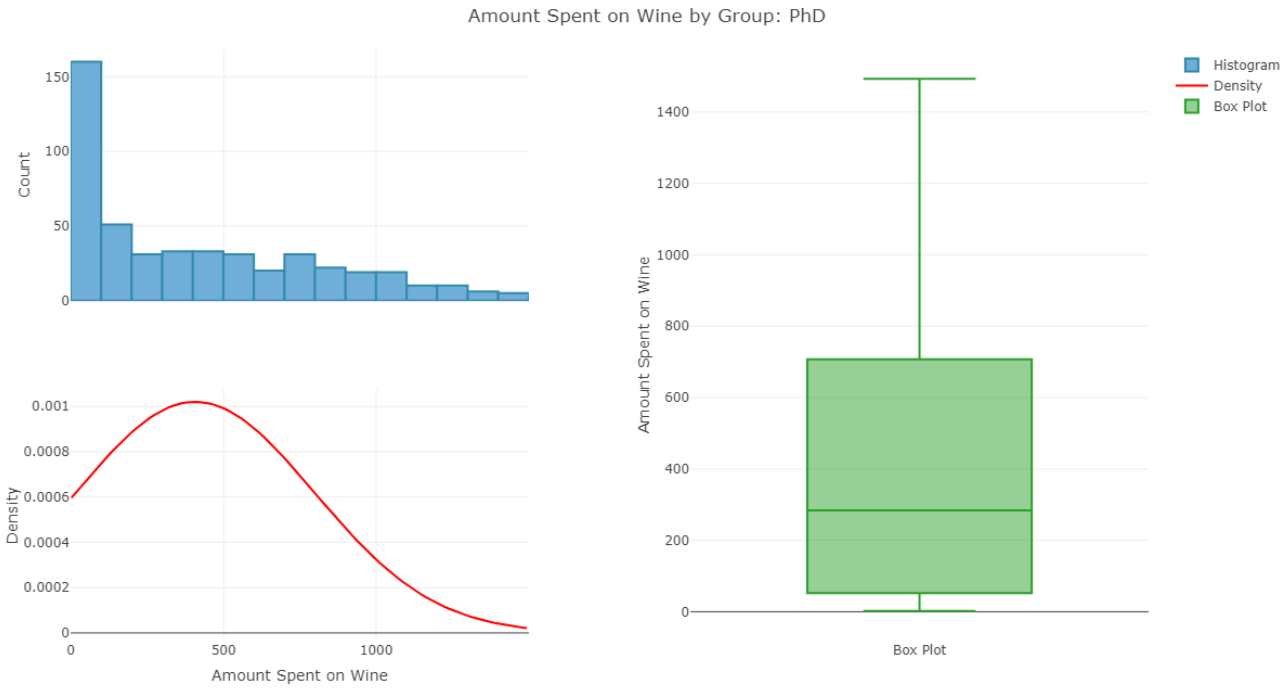
Graduation:



Master:



PhD:



The distributions of all the groups are slightly skewed to the right. None of the groups follows a normal distribution perfectly.

Assumptions of One-Way ANOVA

Independence of samples

There is no evidence to suggest that there is a relationship between the observations in each group. The groups based on level of education are independent.

Therefore, this assumption is met.

Normal distribution of response variable

From the summary plots, we can see that the distribution of the amount spent on wines by all the groups does not follow a perfectly normal distribution. There is a slight right skew in the distribution of the response variable.

Therefore, this assumption is not met.

Since the skew is not extreme, we proceed with the analysis.

Population variance of the response variable for each group

The variance of the response variable for each group is give below.

Group (Education Level)	Variance
2nd Cycle	68916.10
Graduation	95015.92
Master	126821.50
PhD	153044.34

The ratio of the largest variance to the smallest variance is

$$\frac{153044.34}{68916.10} = 2.22$$

The largest value divided by the smallest value is 2.22, which is slightly greater than 2.

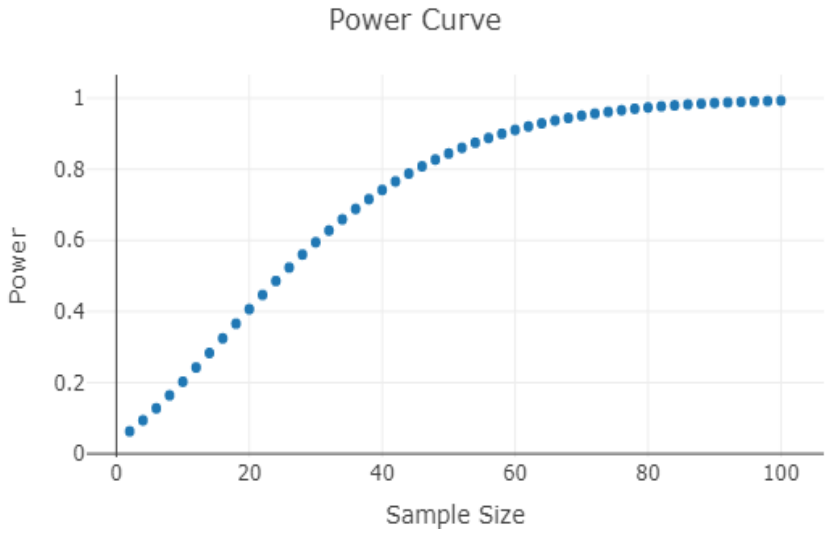
Therefore, this assumption is also not met.

Since the ratio is not significantly larger than 2, we proceed with the analysis.

Power Analysis

The power of the test is calculated using the $power.anova.test()$ function.

The power curve for a significance level of 0.05 is given below.



Group (Education Level)	Number of observations
2nd Cycle	200
Graduation	1116
Master	365
PhD	481

The sample sizes are sufficient for all the groups.

One-Way ANOVA

Global F test for ANOVA:

Step 1:

H_0 – The underlying population means for all four groups are equal.

$$H_0: \mu_{2nd\ Cycle} = \mu_{Graduation} = \mu_{Master} = \mu_{PhD}$$

H_1 – Not all of the underlying population means are equal.

$$H_1: \mu_i \neq \mu_j \text{ for some i and j}$$

$$\alpha = 0.05$$

Step 2

The test statistic is the F statistic with $k - 1$ and $n - k$ degrees of freedom.

Therefore, the test statistic is the F statistic with 3 and 2158 degrees of freedom.

Step 3

The critical F value for $\alpha = 0.05$ and 3 and 2158 degrees of freedom is calculated using R.

$$qf(0.95, 3, 2158)$$

$$F_{3,\ 2158,\ 0.05} = 2.609026$$

Therefore, the critical value is 2.609026

Decision Rule:

Reject H_0 if $F \geq 2.609026$

Otherwise, do not reject H_0

Step 4:

The F statistic is calculated using the model summary of the *aov* function in R:

$$F = 23.46$$

Step 5 – Conclusions:

Since F is greater than the critical value, we reject the null hypothesis.

We have significant evidence at the $\alpha = 0.05$ level to say that there is a difference in the amount spent on wines by the each of the four groups.

Pairwise Comparisons

Step 1:

Null hypotheses:

$$H_{0-1}: \mu_{Graduation} = \mu_{2nd\ Cycle}$$

$$H_{0-2}: \mu_{Master} = \mu_{2nd\ Cycle}$$

$$H_{0-3}: \mu_{PhD} = \mu_{2nd\ Cycle}$$

$$H_{0-4}: \mu_{Master} = \mu_{Graduation}$$

$$H_{0-5}: \mu_{PhD} = \mu_{Graduation}$$

$$H_{0-6}: \mu_{PhD} = \mu_{Master}$$

Alternate hypotheses:

$$H_{1-1}: \mu_{Graduation} \neq \mu_{2nd\ Cycle}$$

$$H_{1-2}: \mu_{Master} \neq \mu_{2nd\ Cycle}$$

$$H_{1-3}: \mu_{PhD} \neq \mu_{2nd\ Cycle}$$

$$H_{1-4}: \mu_{Master} \neq \mu_{Graduation}$$

$$H_{1-5}: \mu_{PhD} \neq \mu_{Graduation}$$

$$H_{1-6}: \mu_{PhD} \neq \mu_{Master}$$

Step 2:

The test statistic is the Studentized Range Distribution with $k = 4$ and $N - k = 2162 - 4 = 2158$ degrees of freedom.

Step 3:

Reject the null hypothesis if the adjusted p value is less than the significance level.

Step 4:

Results of Tukey's Honest Significance test:

Group	p adj
Graduation – 2nd Cycle	0.0055548
Master – 2nd Cycle	0.0000403
PhD – 2nd Cycle	< 1e–7
Master – Graduation	0.0797796
PhD – Graduation	< 1e–7
PhD – Master	0.0073149

Step 5 – Conclusions:

The adjusted p value of Graduation – 2nd Cycle is less than the significance level. We reject the null hypothesis H_{0-1}. Therefore, we have significant evidence to say that there is a difference in the mean amount spent on wines by people with a Graduation education level and people with a 2nd Cycle education level.

The adjusted p value of Master – 2nd Cycle is less than the significance level. We reject the null hypothesis H_{0-2}. Therefore, we have significant evidence to say that there is a difference in the mean amount spent on wines by the two groups.

The adjusted p value of PhD – 2nd Cycle is less than the significance level. We reject the null hypothesis H_{0-3}. Therefore, we have significant evidence to say that there is a difference in the mean amount spent on wines by people with a PhD education level and people with a 2nd Cycle education level.

The adjusted p value of Master – Graduation is greater than the significance level. We fail to reject the null hypothesis H_{0-4}. Therefore, we do not have significant evidence to say that there is a difference in the mean amount spent on wines by the two groups.

The adjusted p value of PhD – Graduation is less than the significance level. We reject the null hypothesis H_{0-5}. Therefore, we have significant evidence to say that there is a difference in the mean amount spent on wines by people with a PhD education level and people with a Graduation education level.

The adjusted p value of PhD – Master is less than the significance level. We reject the null hypothesis H_{0-6}. Therefore, we have significant evidence to say that there is a difference in the mean amount spent on wines by the two groups.

One-Way ANCOVA

Adjusting for Income:

This analysis is different from the one-way ANOVA model in that the for the ANCOVA model, inferences made are based on comparisons made after adjusting for the covariates included in the model. ANOVA is used to compare the means of 2 or more groups, whereas, ANCOVA is used to remove the effect of quantitative metrics before comparing the means of multiple groups.

The results of the model after adjusting for income is given below:

	Sum Sq	Df	F value	p value
Intercept	3775987	1	49.938	2.133e–12
Education	4284903	3	18.890	4.305e–12
Income	76184500	1	1007.561	< 2.2e–16
Residuals	163096859	2157		

Conclusions:

It can be seen that the effect of education level is still significant ($p = 4.305\ e - 12$)

The least square means are given below:

Group	Least Square Mean
2nd Cycle	242
Graduation	287
Master	334
PhD	384

The least square means are the group means after controlling for the covariates.

The mean amount spent on wines by people with a 2nd Cycle education level is 242.

The mean amount spent on wines by people with a Graduation education level is 287.

The mean amount spent on wines by people with a Master education level is 334.

The mean amount spent on wines by people with a PhD education level is 384.