**CIDM 6355 Exam 2 Part 2 Submission**

**(75 points in total; due 11:59 pm CDT, April 22nd 2024)**

Requirements: This exam is open book, open slides, and open notes. However, this is an individual exam, so you are not allowed to collaborate nor discuss with anyone else before the due time of the exam. Any question about the exam should be addressed to the instructor. You are required to follow the instruction to complete all the questions and deliverables. You are not allowed to share your RM processes, R scripts, screenshots, or answers with other students or parties; otherwise, such a behavior will be reported to the university authority. In addition, it is your responsibility to make your answers meet the required format; otherwise, you might lose points because of wrong format.

Please read, understand, and comply with these requirements in this exam by typing your name as below.

**Name: Mariam Adegbindin**

1. Step 2.3 Take a screenshot of the centroid table with date and time (Screenshot 1) and briefly describe each cluster. [8 points: 4 pts for your screenshot and 4 pts for your description]

A screenshot of a computer

Description automatically generated

**Cluster 0:** Wines in this cluster are characterized by moderate levels of fixed acidity, indicating a balanced acidity level that is neither too high nor too low. They also have relatively low residual sugar content, suggesting a drier profile. Total sulfur dioxide levels are moderate, indicating some degree of oxidative protection. The wines have low density, potentially indicating a lighter body. pH levels are moderate, contributing to the overall balance of acidity. Alcohol content is also moderate, providing a pleasant level of alcohol warmth without being too overpowering. Overall, wines in this cluster may offer a well-balanced and approachable drinking experience.

**Cluster 1:** This cluster represents wines with low to moderate levels of fixed acidity, suggesting a range of acidity levels across the wines. They exhibit high residual sugar content, indicating a sweeter profile that may appeal to those with a preference for sweet wines. Total sulfur dioxide levels are high, potentially indicating greater oxidative protection or preservative levels. Density is moderate to high, suggesting a fuller body and potentially richer mouthfeel. pH levels are moderate, contributing to the overall balance of acidity and sweetness. Alcohol content is relatively low, which may contribute to a lighter and less intense drinking experience.

**Cluster 2:** Wines in this cluster are characterized by low levels of fixed acidity, indicating a milder acidity profile. They also have low residual sugar content, suggesting a drier taste profile similar to Cluster 0. Total sulfur dioxide levels are moderate, providing some degree of oxidative protection. Density is low, indicating a lighter body and potentially more delicate flavors. pH levels are high, indicating a less acidic taste profile. Alcohol content is moderate, contributing to a balanced and harmonious drinking experience. Overall, wines in this cluster may offer a refreshing and easy-drinking style with subtle flavors and a smooth finish.

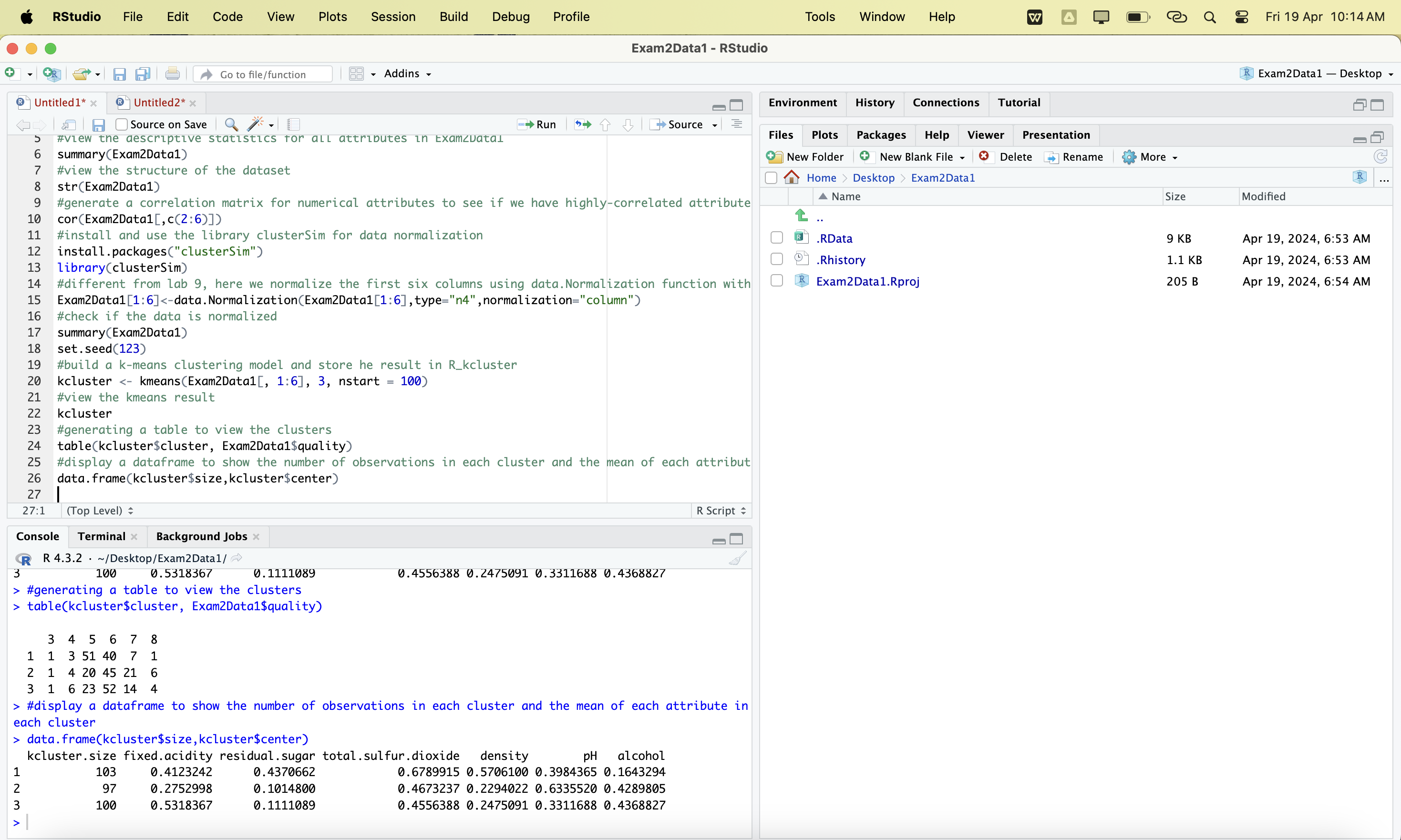
1. Step 2.6 Take a screenshot of the ANOVA table with date and time (Screenshot 2) and briefly describe your conclusion. Your conclusion must be based on both Steps 2.4 and 2.5. [8 points: 4 pts for your screenshot and 4 pts for your description]

A screenshot of a computer

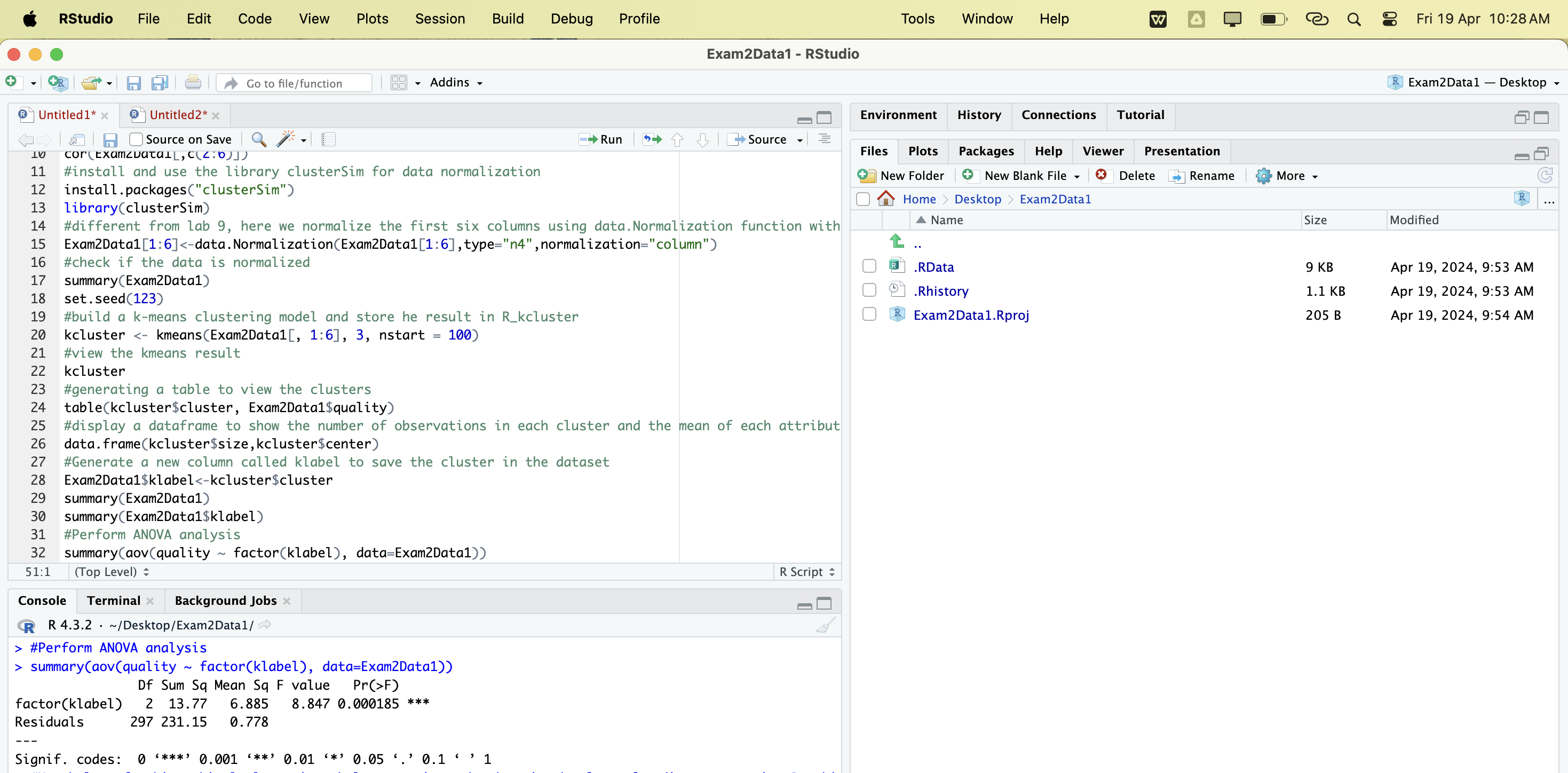
Description automatically generated

It can be observed that Cluster 0 exhibits the lowest average quality, whereas Cluster 2 demonstrates the highest, with Cluster 1 occupying an intermediate position in the bar chart. Based on the obtained p-value being less than 0.05 (p < 0.05), it can be inferred that the probability of obtaining the observed results, or more extreme results, under the assumption of the null hypothesis being true, is exceedingly low. This implies that the observed differences in average quality ratings among the three clusters are highly unlikely.

1. Step 3.3 Take a screenshot of your output (cluster size and centroids) with date and time (Screenshot 3) [4 points]



1. Step 3.6 Take a screenshot of the ANOVA table with date and time (Screenshot 4) and briefly describe your conclusion based on the ANOVA table. [8 points: 4 pts for your screenshot and 4 pts for your description]



The p-value, calculated at 0.000185according to the ANOVA table, suggests a notable difference among the compared groups. This value is substantially lower than the conventional alpha level of 0.05, indicating statistical significance in average quality ratings of the wine.

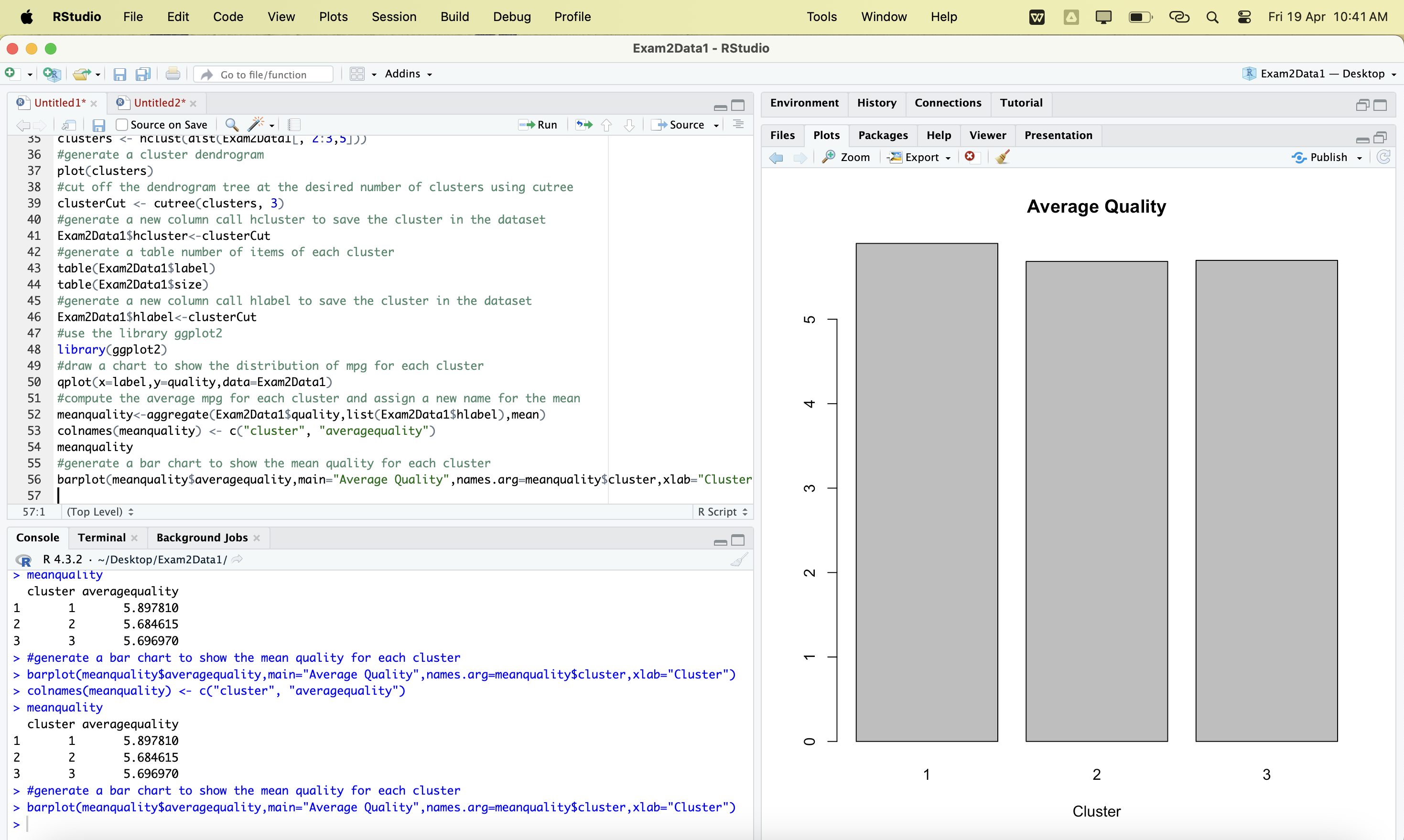
1. Step 4.3 Take a screenshot of the bar chart with date and time (Screenshot 5) and briefly describe your conclusion. Your conclusion must include each cluster’s size and their average quality ratings. [8 points: 4 pts for your screenshot and 4 pts for your description]

A screenshot of a computer

Description automatically generated

The analysis of the wine clusters reveals distinct patterns in quality ratings and cluster sizes. Cluster 1 emerges as the standout group with the highest average quality rating of 6.36, indicating wines of superior quality. In contrast, Cluster 2 exhibits the lowest average quality at 5.71, suggesting wines of relatively lower quality. Cluster 0, being the largest cluster with 229 wines, presents a middling average quality rating of 5.66. These findings suggest that certain attributes or characteristics differentiate wine clusters, influencing their perceived quality.

1. Step 5.5 Take a screenshot of the bar chart with date and time (Screenshot 6) and briefly describe your conclusion. Your conclusion must include each cluster’s size (Step 5.3) and their average quality ratings. [8 points: 4 pts for your screenshot and 4 pts for your description]



Cluster 1 has the highest average quality rating at 5.90, followed by Cluster 3 with an average quality rating of 5.70, and Cluster 2 with an average quality rating of 5.69. The difference in average quality ratings between the clusters is relatively small, with Cluster 1 showing a slightly higher rating. While Cluster 3 and Cluster 2 show relatively similar average quality ratings, the slightly higher rating of 5.90 for Cluster 3 suggests that it may contain wines of marginally better quality compared to Cluster 2.

1. Step 6.2 Please compute the match rate of each pair of models and attach the corresponding screenshot below. You must show how your match rate is computed; your screenshot (e.g., a PivotTable) must show how clusters from each model matched with each other. Your screenshots do not have to show date and time. [12 pts: 2 pts. for each match rate and 2 pts for each screenshot].

In order to calculate the match rate, please explain how it is computed, using the following example provided in Exam 2 Part 2 Instruction as a reference. Failure to do so will result in a zero point on the match rate you provide. Additionally, you must provide a screenshot clearly showcasing how each pair of clusters is matched, similar to the example provided. Failure to provide this will also result in a zero point on each the screenshot you provide.

|  |  |  |
| --- | --- | --- |
| Model Pair | Match Rate | A screenshot to support your match rate |
| Models 1 & 2 | (100+103+97) /300 =100% |  |
| Models 1 & 3 | (93+43+21) /300  = 52.3% |  |
| Models 3 & 4 | (229+50+21) /300  = 100% |  |

1. Step 6.3 Why some pairs of clustering models above do not generate the same clustering results? [4 bonus pts].

K-Means assigns each data point to the nearest centroid, which can lead to different results based on initial centroids while agglomerative Clustering builds a hierarchy of clusters by merging data points step by step. Another reason the clustering models do not generate the same clustering result is both K-Means and Agglomerative Clustering are nondeterministic. They rely on random initialization (e.g., initial centroids or linkage criteria). Running the same algorithm multiple times with the same input data may yield different results due to different initializations.

1. Step 7.2 Take a screenshot of your RM process (if you are using the operator “Work on Subset”, please take two separate screenshots: one illustrating the overall process and the other depicting the subprocess nested within the "Work on Subset" operator). Take a screenshot of your Decision tree model (the Graph view) and Neural Network model (the Neural Net view). [12 points: 4 pts for each screenshot and three screenshots in total; if you have four screenshots, each would be worth 3 pts.]

A screenshot of a computer

Description automatically generated

*Decision tree model (the Graph view)*

*A screenshot of a computer

Description automatically generated*

*Neural Network Model (the Neural Net view)*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

1. Step 7.3 Discuss the benefits of such a two-step approach (Clustering and then Classification) and provide a practical example that demonstrates this two-step approach [7 points: 3 pts for the benefits and 4 pts for a practical example; you need to describe how the two-step approach is used in the example you provided].

Benefits

Clustering helps identify underlying patterns or structures in the data that might not be immediately apparent. By first partitioning the data into meaningful clusters, classification models can be trained on more homogeneous subsets, potentially leading to improved performance. In cases where the dataset is large and complex, clustering can serve as a preprocessing step to reduce the computational burden of classification algorithms. By dividing the data into clusters, each classification model can be trained on smaller, more manageable subsets.

Example

Using an example where a company wants to identify customer segments and then classify these segments based on their likelihood to purchase a new product.

**Clustering**: The company could apply a clustering algorithm such as k-means to group customers into distinct segments based on features like demographic information, purchase history, and engagement metrics. This clustering process would partition the customers into clusters such as "high-spending frequent buyers," "budget-conscious occasional buyers," and "inactive non-buyers."

**Classification**: Once the clusters are identified, the company can train separate classification models for each cluster to predict the likelihood of purchasing the new product. For example, a logistic regression model could be trained on the "high-spending frequent buyers" cluster to predict whether a customer in this segment is likely to purchase the new product within a given time frame. Similarly, models could be trained for the other clusters.