**Project Requirements**

**Goal:** Demonstrate proficiency in data handling and decision-making.

**Synopsis:**

BurgundySip is a successful wine cellar/store that promotes high-quality wines from all over the world. BurgundySip management recently decided to use data analytics to streamline their pricing strategy. You have been hired as part of a team of analysts to conduct a pilot exploratory (but still prescriptive) study to investigate data relationships, trends, and patterns in price as well as other variables.

You have been given a subset of data from their vast data warehouse. The dataset contains the following variables:

* SN: Serial number of the sample.
* NAME: Winery name
* WINE: Name of the wine
* YR: Year in which the grapes were harvested – indicator of the AGE of the wine.
* RT: Average rating given to the wine by the test users [from 1-5]
* NUMR: Number of testers that reviewed the wine
* REG: Region of the wine
* PR: Price in euros [€]
* TP: Wine variety
* BD: Body score, defined as the richness and weight of the wine in your mouth [from 1-5]
* ACD: Acidity score, defined as wine's “pucker” or tartness; it's what makes a wine refreshing and your tongue salivate and want another sip [from 1-5]
* RSG: residual sugar level of the wine [from 0 -16]
* AL: Alcohol percentage of the wine.
* DN: The typical density or specific gravity of the wine is generally between 1.080 and 1.090.

**Your Task:**

1. Identify a series of analytical objectives. These must include:
   1. At least 5 descriptive (statistical) questions.
   2. At least 2 predictive objectives (2 regression models).
   3. At least 2 cluster-based learning objectives.
2. Program in R to achieve these objectives by following all the steps of the data handling process ***(Deliverable 2)***. Each group needs to submit ONE R code file. The file must be professionally arranged, and each code segment must be clearly explained by using comments.
3. Design a PowerPoint presentation with “professional speaker notes” as per the requirements below ***(Deliverable 1)***. ***These notes should explain each slide in detail*** (You do not have to read out all the notes during the presentation). This is a technical presentation; Therefore, it must be professionally presented with adequate technical details. Each group must record Each group must record their presentation using any collaboration tool and submit it along with the PPT file.
4. Submit your PowerPoint file, video presentation, R code file, and individual reflection file before 11.59 PM on December 18, 2022.
5. When submitting the video, please upload it in a shared folder and submit only the link to the video.

**Deliverables 1 and 2:**

The following sections must be clearly reflected in both deliverables 1 and 2. [Please refer to the attached rubrics for score breakdown]

1. Introduction
   1. Briefly introduce the case, data handling process, and presentation structure.
2. Analytical objectives. These must include:
   1. At least 5 descriptive (statistical) questions.
   2. At least 2 predictive objectives (2 regression models).
   3. At least 2 cluster-based learning objectives.
3. Assumptions
   1. List all the assumptions you had to make to complete the analysis here.
4. Data Analysis Process
   1. Briefly describe the analysis process (include a process map).
   2. You may refer to the subsequent sections. This section may outline the underlying theories. Please do not include any implementation steps or findings in this section.
5. Data Sanity Checks
   1. Importing Data.
   2. Variable Analysis
      1. Perform a variable summary.
      2. Perform an initial visualization if required.
      3. Check for the structural integrity of the dataset.
      4. If any variable is not conforming, you may consider casting/conversion.
      5. Recognize and treat categorical variables.
   3. Treating Duplicate Entries
      1. Describe the process/ strategy.
      2. Describe the findings/results.
   4. Treating Outliers and Missing Values
      1. Describe the process/ strategy.
      2. Describe the findings/results.
6. Descriptive Analysis and Findings: Data Summarization, Visualization, Statistical Analysis, and Calculations
   1. Describe how the calculations were performed and why they had to be performed.
   2. Describe the use of visual artifacts in the analysis.
   3. Explain the findings.
7. Predictive Analysis and Findings
   1. Describe how regression as a predictive analysis technique was used to achieve the objectives.
   2. Describe the use of visual artifacts in the analysis.
   3. Explain the findings.
8. Cluster-based Analysis and Findings
   1. Describe clustering was used to discover classes of market segments/ achieve the set objectives.
   2. Describe the use of visual artifacts in the analysis.
   3. Explain the findings.
9. References: Please follow the standard Microsoft APA referencing style.

**Marking Rubric: Deliverable 2 – R Code**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question | 0-20% | 20-50% | 50-80% | 80-100% | Final Marks |
| Sections | Invalid or incorrect answer. The code doesn’t run, and the steps are incorrect. | The code works but the intermediate steps are not compatible with the answer. A correct answer with a poor approach. Some of the steps are correct, although the code doesn’t work. | The code works and most of the steps meet the question’s expectations. Some comments are available. Most of the steps are correct, although the code doesn’t work. | The code works and all the steps meet the question’s expectations. The code is well commented and well-structured. | **Marks Out of 80** |
| Importing Data | 0-2 | 2-3 | 3-4 | 4-5 |  |
| Variable Analysis | 0-2 | 2-5 | 5-8 | 8-10 |  |
| Treating Duplicate Entries | 0-2 | 2-5 | 5-8 | 8-10 |  |
| Treating Missing Values | 0-2 | 2-5 | 5-8 | 8-10 |  |
| Descriptive Analysis | 0-2 | 2-5 | 5-8 | 8-10 |  |
| Predictive Analysis and Visualization | 0-3 | 3-7 | 7-12 | 12-15 |  |
| Cluster Analysis and Visualization | 0-2 | 2-5 | 5-8 | 8-10 |  |
| R code file quality | 0-2: Poor quality | 2-5: Structure is maintained. Some code is explained using comments. | 5-8: Structure is maintained. Most of the important code sections are explained using comments. | 8-10: Structure is maintained. All important code sections are explained using comments. |  |
| Total | | | | |  |