| Student | Li Chenyang | 80 |
|---|---|---|

## Part 1

| | Available | Marks |
|---|---|---|
| The file *roller_coasters.csv* contains data on a number of roller coasters from all around the world. This shows the name, type, material, manufacturer, location and current status for each roller coaster and may also show data on the height, length, speed, and number of inversions achieved by each roller coaster. Load this data into a data frame *coasterData*, ensuring that the headers are imported and that the type, material, manufacturer, and current status are imported as factors. Your script should assume that the data file is stored in the same folder as the script file. Display the resulting data frame. | 5 | 5 |
| Create a new data frame *highestData* that shows the name, height, and location of the highest operating roller coaster of each type of roller coaster. Display this data frame. | 10 | 8 |
| Include a function *displayComparison()* that will take the material as a parameter and will produce a bar chart showing the speed of the top ten fastest operating rollercoasters of that material. Display this using wooden roller coasters. | 15 | 14 |
| Include a function *correlateHeightSpeed()* that generate a data frame *heightSpeed* that shows the mean height, mean speed, and coefficient of correlation of height and speed for each type of roller coaster and each material of roller coaster. These calculations should be based on records in *coasterData* that contain no NA fields. | 15 | 10 |
| Commenting, white space and code presentation | 5 | 5 |
| *TOTAL* | | 42 |

Comment

Q1: Good

Q2: Good. Good use of aggregate().

Q3: Good. Good idea to rotate the text to make it clearer.

Q4: The calculation of correlations is good, but the output contains too many NA values. It would have been better to filter combinations of material and seating types to remove those that don't occur together.

Good commenting and use of white space, but remember to add your name and student ID to the top of the script.

## Part 2

| | Available | Marks |
|---|---|---|
| The file *Covid.xlsx* contains data from Our World In Data (https://ourworldindata.org/covid-vaccinations) which shows data collected from official government websites worldwide on the progression of the Covid-19 pandemic. It contains 2 sheets, one showing the data for China, and the other showing the data for India. Load this data into a data frame *covidData*, ensuring that the headers are imported. Your script should assume that the data file is stored in the same folder as the script file. | 5 | 5 |
| Plot the number of cases and number of deaths by month for both countries on the same appropriately-labelled graph. | 10 | 8 |
| Create a data frame *CovidSummary* that shows descriptive statistics for the total number of new cases in China and India for March 2020, September 2020, March 2021, September 2021, March 2022 and September 2022. This should contain the following measures: mean, standard deviation, median, max, min, and range. | 15 | 5 |
| Perform a t-test to identify whether there is a significant difference in the numbers of new cases in India between September 2021 and September 2022. You should state your experimental hypotheses, the result of the test, and whether you accept or reject the null hypothesis in your comments. | 5 | 5 |
| Perform an ANOVA test to identify if there are statistically significant differences in the numbers of new cases in China in March 2020, September 2020, March 2021, September 2021, March 2022 and September 2022. If such a difference is found, perform the appropriate ad-hoc tests to identify where the differences lie. You should state your experimental hypotheses, the result of the test, and whether you accept or reject the null hypothesis in your comments, and should explain what you are doing and why you are doing this in the comments. | 10 | 10 |
| Commenting and use of white space | 5 | 5 |
| *TOTAL* | | 38 |

Comments

Q1: Good.

Q2: Good. Try to avoid using exponential notation on axes.

Q3: You've got the statistics for the overall data, not the time- and country-limited data.

Q4: Good.

Q5: Good.

Good commenting and use of white space.