

Student	Madhusudan Ghumare	31
---------	--------------------	----

Part 1

	Available	Marks
The file <i>roller_coasters.csv</i> 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 <i>coasterData</i> , 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	2
Create a new data frame <i>highestData</i> that shows the name, height, and location of the highest operating roller coaster of each type of roller coaster. Display this data frame.	10	2
Include a function <i>displayComparison()</i> 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	6
Include a function <i>correlateHeightSpeed()</i> that generate a data frame <i>heightSpeed</i> 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 <i>coasterData</i> that contain no NA fields.	15	2
Commenting, white space and code presentation	5	5
TOTAL		17
Comment Q1: You need to put the data file in the same location as the script file. Type, material, manufacturer and current status should be converted to factors. Q2: You've got all of the data, not data for the highest coasters. Q3: You've correctly created a function to draw a barplot, but you haven't filtered to get the top ten fastest coasters. Q4: You have not created a function. You have manually created a data frame <i>heightSpeed</i> but you need to create that automatically to show the correlations between height and speed for all material types. Good commenting and use of white space.		

Part 2

	Available	Marks
The file <i>Covid.xlsx</i> 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 <i>covidData</i> , 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	2
Plot the number of cases and number of deaths by month for both countries on the same appropriately-labelled graph.	10	2
Create a data frame <i>CovidSummary</i> 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	4
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	1
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	1
Commenting and use of white space	5	4
TOTAL		14
Comments Q1: Don't give the full pathname if your code is going to be run on a machine that you don't know the directory structure or it won't run. The data should be stored in a dataframe called <i>covidData</i>. Q2: A set of boxplots is not an appropriate method of displaying this data. A bar plot or line graph would be suitable. None of the plots make it clear which are India and which are China. Q3: You need to write code to calculate the number of new cases, not just type them into a vector. The code should then have calculated the descriptive statistics for each month and each country. Q4: You are trying to perform t-tests on data that is not the data required for the problem. You need to state your hypotheses, extract the appropriate data from the dataset, run the test, then interpret the results. Q5: You are trying to perform ANOVA test on data that is not the required data for the problem. You need to state your hypotheses, extract the appropriate data from the dataset, run the test, run any appropriate post-hoc tests then interpret the results. Commenting and use of white space are good.		