2021 S1 DATA1001 Main Exam

Note to Markers: Below is the overall scheme. You may find some other sensible alternatives which satisfy the questions - however, the exam is what differentiates between students (based on usually high project marks), so the answer must be precise and concise, showing careful statistical thinking. No marks for just 'guessing'.

Question	Sample Answer	Mark	Level
21(i)	2 confounders specific to teenage mental health. Eg Peer pressure, parent's diet, maturity to describe personal situation. Note: Can be listed, without context. Can't be unrelated to context. Can be related to bias. ½ mark each.	1	Cr
21(ii)	Method for data collection specific to teenage mental health and how it would be implemented allowing "large" sample. Eg: An iphone app that teenagers in mental health treatment are asked to use. 1 mark for sensible suggestion. ½ mark for good attempt (eg "online survey").	1	P
21(iii)	Assessment of whether RCT is possible, with evidence, in context! Eg: It is possible, but fairly difficult, given ethical issues with giving and with-holding potentially live saving treatment to someone with suicidal tendencies. Eg: It is possible, if the research question involves say social media, so the random allocation does not involve mental health. ½ mark if describes RCT. If (iii) linked to (i), then 1 mark.	1	CR
22(i)	One sensible limitation + one possible solution. Eg One limitation is that the data does not include many countries. A solution would be to increase the data collection/reporting for those countries or be conscious to state all conclusions only for those countries represented in the data. Eg Data wrangling needed to get exact location in terms of street address. 1 mark each.	2	P

22(ii)	Amador is in Wien Austria with zipcode 1190 and serves Creative Cuisine at a very high price. (Looking at the url, also gives more specific info. We are assuming $$$$ means high price.) 1 mark for drawing at least 2 characteristics from the head() code.	1	P
23	3 careful sensible comments, with any assumptions. Eg Assumption: we assume that more $ signs means higher price in the Price variable. The most expensive food is Creative, Innovative, Korean and Sushi [need 3+] or uniquely Innovative and Sushi. The ‘cheapest’ food (relatively for 3 star restaurants) is Cantonese and Chinese. The most common type of cuisine is Contemporary. 1 mark for each comment. 1 mark available for assumptions, if only 2 comments awarded.	3	CR
24(i)	1 mark for a correct option (ie a place close to "green") - not Oslo, Seoul, Stockholm.	1	P/CR

24(ii)	1 mark for documentation (use of #)
	1 mark for correct ggplot
	1 mark for geom_bar() or geom_col() [though 2nd one should have reshaped data]
	1 mark for aes(fill=cuisine))
	1 mark for title (or labelling x axis or removed legend title)
	
	# do ggplot, with perpendicular labels and title
	ggplot(stars3, aes(x = city)) + geom_bar(aes(fill=cuisine)) + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) + ggtitle("Michelin Guide 3 star: Price over cuisine")
	
	Note: students may link (i) and (ii) and instead do a plot for Contemporary. Can give the 5 marks.
	Possible variation: ggplot(stars3, aes(y = city))
		5	HD

25(i)	Sensible answer that doesn't just rely on R output. Answer includes a choice of classification and some justification (based on domain knowledge, eg rating scale). Eg R classifies the variables as ‘int’. Whether this is appropriate depends on how the rating scale from 0 to 30 is constructed. It could be quantitative, but it could be qualitative ("rating" scale). Not enough to say 'quantitative' as numbers, or R classifies as 'int'. 1 mark for any attempt at an explanation that considers the nature of the variables. 0 if only classification.	1	CR
25(ii)	nycfood$East=factor(nycfood$East) or as.character() or as.logical(nycfood$East) or any sensible alternative. ie Not as.numeric()	1	P
26(i)	4 variables: Food (x axis), Price (y axis), East (shape of points), Service (colour of points) ½ mark for just recognising "4".	1	P
26(ii)	3 precise, concise comments on context: 1 mark each. Eg Generally higher price is associated with higher ratings of food and service. Position relative to East Avenue results in range of food/service ratings. However, there is much variation with 1 restaurant having poor service and West of 5th Avenue, but having high rated food and high price.	3	CR

27	1 mark for each of HATPC. S: Let alpha = 0.05. Ho: There is not a linear relationship between Price and Food rating vs H1: There is a linear relationship between Price and Food rating. ½ mark if no mention of 'linear'. A The residuals should be independent, normal, with constant variance (homoscedasticity). Check: Residual Plot looks random. [can be combined with statement of assumptions] Not enough to just state assumptions, but OK to go straight to arguing that the assumptions are valid. T test statistic is 10.371 (accept F statistic of 107.6 - same p-value) P p-value is very small C Given p-value is so small (eg relative to 0.05), we reject Ho, which means we conclude that there appears to be a linear relationship between Price and Food. ½ mark if C doesn't reference p-value.	5	CR/D
28(i)	Any good attempt that at least mentions processes of Michelin-starred chefs + the theoretical model.	1	D
28(ii)	A sample of 12 Michelin-starred chefs awarded one, two or the maximum of three stars from Germany were interviewed. ½ mark for partial answer.	1	P
28(iii)	2 sensible comments: 1 mark each. Eg Germany not neccesarily transferable to NYC or sample size. Eg Importance of employees. Eg Place of new food creation in top tier restaurants.	2	HD
		30	