2024 RMHI/ARMP Problem Set 2

Hello everyone! This is the description for the assignment, which is due on Canvas on Monday May 13, 2024 before 8:00am Melbourne time. You'll need to submit a Word-knitted version of the completed R Markdown file found in this zip file, according to the following instructions:

1. Rename the document called pset2.Rmd as studentID-pset2.Rmd. (Replace studentID with your student ID number). This is your R Markdown file, where you'll be putting your code and answers.

2. Replace "Your name and ID goes here" in the header of the R Markdown file with your name and student ID. (Keep the quotes or it won't knit properly).

3. While we encourage collaboration in tutorials and learning in general, *you should not be collaborating with anybody AT ALL for this assignment.* **That means sharing code privately or publicly; even talking in the abstract about problems will effectively be collusion.** You should be completing it independently, with no help from any other person in any capacity. Of course, as always, you are free to use any of the resources from the class to help you, and you're also free to google or look anything up that you like (as long as you aren't asking anybody, including discussion boards or AIs, questions related to this assignment). Note that we do look at places like chegg and will follow up if anything from this problem set is posted there.

4. Plagiarism check is enabled and you can check the similarity report on your submission. In previous years we have found people who tried to cheat, so please don't risk it! That said, understand that we will not be naively looking at the overall % figure: with this sort of assignment a certain amount of overlap is inevitable, so don't worry if you get what looks like a high % score as long as you know you didn't plagiarise or collude. With this sort of assessment, that % overlap is higher than essays and the like. We will be using the plagiarism check for the parts of the assignment where we'd expect some variability, and to give a general sense of the overall gestalt.

5. Complete all of the problems below in the R Markdown document. *Do not remove any of the arguments to the code chunks, like the names of the code chunks or where it says message=FALSE or whatever.* If a problem asks you to display a tibble or variable so it shows up in the knitted version, **make sure that you do** as the marker cannot evaluate it without seeing it, and if they can't see it then they won't be able to award you points for it! Remember that to display a tibble (or any variable) you just type its name on a line of its own within the R chunk.

6. As much as possible, **questions do not build on each other** (often even parts of questions). That means that if, say, you can't get Q5 then you can still get Q6. Try to do all of them.

7. **Go for partial credit!** Many of these questions have some form of partial credit possible. If it is asking for some R code, break down the problem into pieces. Even if you can only do some of the pieces, or do them part of the way, that will be worth something. [Note that there is no question-by-question rubric available because designing one would mean giving away the answers. In general we will give full credit for responses that correctly address all of the parts of the question.] Short answer questions (SAQs) can also be given partial credit and are generally asking for some thoughtful interpretation. If it is based on a previous graph or test you've done, if you did the first part wrong but discuss it well, you can still get most or all points for the SAQ part. If your code does not run but you want to include it for possible partial credit, just comment it out (using the # sign) so that it shows up in the knitted document but R does not try to run it. If some of your commented-out code is correct and some isn't, we will not give you credit for the correct code; put the thing in there that you think is the closest to the correct answer, don't just include everything.

8. We don't care much what decimal place you round answers to and you will not lose credit for decimal places unless you round so much that your answer is impossible to discern (e.g., don't round p-values to the nearest integer!). Similarly, you will not lose points for trivial presentation things like using parentheses instead of commas around statistical references, as long it's clear. That said, for those who want a guideline, I'll suggest that you follow APA format or round p-values to three decimal places, degrees of freedom to one, and test statistics and probabilities to two.

9. Some questions specify a word count. In that case you need to either calculate it from the knitted document or type up your answer in Word[1] and then cut and paste it into the R Markdown file. (Please put your answer in between the word ANSWER and [Word count: XX]; those two bits do not count towards your word count). The word counts specified in each question are designed to give you a guideline about the maximum number of words you should need answer completely and correctly. So don't feel like you must use all of the words; if you can answer it fully with less, that's fine. In fact, the total word count for the solution set we wrote up is around 1320, so it's possible to fully answer the questions while going substantially under the word limit. That said, it is okay to go over the word limit for individual questions as long as the total word count for all of the questions combined is fewer than 1760 words (i.e., fewer than 1600+10%, with the standard penalty if it is 1600+10% or over. See the student manual for details on word count penalties).

10. There is no word count for code chunks. Word count only applies to the short answer questions that have [Word count: N] attached and suggested word counts given. **Remember to report** your total word count for the assignment as a whole at the top of the document. Your total word count is the sum of the word counts where those are reported.

11. You'll be turning in the knitted output of your R Markdown file. We prefer that you knit to Word but if you can't get Word to knit then html is okay. In the worst case, you can turn in the completed Rmd file. **We highly, *highly* recommend that you knit as you go**: (a) knitting can identify problems in your code that you would have otherwise missed; and (b) you do not want to get close to the deadline and think you're done only to find that you're having trouble knitting. Save yourself the panic and knit often.

12. Similarly, **you can turn in the assignment multiple times before the deadline**, so we strongly encourage you to submit it even before it's perfectly polished. That will save you last-minute panic or computer issues. Also, take a screenshot for proof of having turned it in just in case you need it. If you run into last-minute computer issues and can't even succeed in uploading an Rmd, *email us (rmhi-armp@unimelb.edu.au) your assignment as soon as possible* to demonstrate that it was done at that time. We cannot make promises about whether you will receive any late penalties if you do this, but if you don't, you almost certainly will get penalised because we have no way to know if the problems were genuine.

<div align="center">*** IMPORTANT NOTE ***</div>

Making figures is an important part of doing your data analysis, and in many of these problems creating the appropriate figure will help you understand what is going on with your data. Deciding out *what* figure is necessary and *when* is also a key skill.

What this means is that **even if a problem does not explicitly say you should make a figure or what it should look like, you may benefit from making one anyway!** As a result, all of the code chunks include options for figure sizes regardless of whether one is necessary; you can leave the code chunk as is even if you don't make a figure. If a question doesn't tell you to make a figure, you will always be evaluated on the correctness of your analysis only, and if you can understand and explain[2] the data without creating a figure that is fine. However, *much of the time* a figure will help understand and explain the data, and you will never be penalized for making a figure that you don't need (unless you interpret it wrong or something). So if you do not understand something about your data or analysis, make figure(s) or tables! If you do make one, you won't be evaluated on its beauty but if it is unclear then you will not benefit from making it.

---

[1] We know different software calculates word count in slightly different ways, so we are using Word as the standard.

[2] **If you are asked to include descriptive statistics, showing a figure with the relevant information is sufficient;** as long as the information is clear in the figure (and the marker can see the figure) then you don't need to also list all of the descriptive statistics in the text. If the figure contains error bars, the error bars can indicate either standard error or standard deviation; unless one of these is explicitly asked for by the question, either is fine as long as you indicate which it is in the subtitle or text. If you calculate the descriptive statistics in your code but there is no figure then you need to report them in the text.

<p align="center">**The story of LFB and Foxy**</p>

* * *

For this assignment, we're going to go back to meet up with LFB and Foxy and hear their story. As you'll recall, a few weeks ago they, Doggie, and Flopsy went on a mission to Otherland to steal some of their data. The mission was successful but LFB and Foxy went missing! In your assignment we get to see what happened to them.

* * *

LFB and Foxy are acting as lookout as Doggie and Flopsy enter the building. Standing on one side of the building, LFB is squinting through the darkness trying to see when she hears a rustle. Then another one, then another one, coming ever closer. Not wanting to raise the alarm prematurely, LFB holds still, but when she hears another rustle only meters from where she is, she whistles, giving the signal.

Foxy dashes around the building as quickly as she can, just in time to see the bushes near LFB part. She catches a hasty glimpse as three very large shapes -- bears? dogs? something else? -- jump out at LFB. Startled, LFB whistles as loud as she can, but it is abruptly cut off by one of the animals covering her mouth.

Forgetting her normal shyness, Foxy shouts "stop!" as loud as she can and charges at the two creatures. She growls at them, surprising herself, and they turn. She trembles: one of them is the largest bear she has ever seen, but it's too late to go back now. She growls again, and then the bear rushes at her and hits her and she is knocked unconscious.

After a short and frightening journey through the dark, LFB and Foxy are put into a small, bright room. LFB is relieved to see Foxy start to stir after a moment — she hasn't been hit that hard — and the two of them cuddle together in fear. The room they are in looks like a library, but it has very large chairs and books. LFB has to jump just to get down from the high sofa and reach the door handle. It is locked. There is a small window in the door, too high for either of them to reach.

Eventually they notice that different people keep looking through the window; it appears that the people are taking turns watching them. The monitoring shifts are shared between three people: a bear who is so large that only one eye fits in the window at a time, a horse (or possibly unicorn?) of some sort, and what might be a hippo or rhinoceros. Not having anything better to do, Foxy and LFB keep track of how long each person spends watching them before the next person takes over.

Q1 [20% of total mark]

After 60 shifts (20 from each person), LFB wonders if there is a difference between the three people in the amount of time they spend watching on average per shift.

The tibble ds, which has been loaded for you in the R Markdown document and is described in Table 1, contains the information about the shifts that LFB and Foxy have gleaned.

Variable	Description	Table 1: variable description for tibble ds
shift	The number of shift during which one of the animals watched Foxy and LFB through the window in the door. There are 60 shifts total, and each person watched them for 20 shifts.	
person	The person watching during that shift. Since they don't know their names, we will use the nicknames LFB and Foxy came up with for them: ScaryBear, BigHippo, and QuietHorse.	
length	Amount of time in seconds that person spent watching them on that shift.	

Q1a. Create a bar plot showing the mean length of time spent watching per shift: there should be three bars, one corresponding to each person. Each bar should be outlined in black and have error bars corresponding to the standard error. The colour of each bar should be semi-transparent and be different for each person, using a sensible palette of your choice. Individual data points should be included and follow the same colour scheme as the corresponding bar, and both bar and individual data points should be visible. There should be no legend. Title and label the axes appropriately (you don't need a subtitle). Make sure your figure shows in the knitted document.

Q1b. There are two main assumptions you need to test before implementing the analysis in Q1c below that addresses LFB's question. Evaluate each of these assumptions and report the results. (It does not matter what order you test them in). For each assumption, be sure to explicitly state what assumption you are testing, the name of the test you did to evaluate it, the results including the stats reference, whether the assumption was violated, and how you could tell.

[Suggested word count: 90]

Q1c. Based on your results in Q1b, run the appropriate statistical test to evaluate whether there is a significant difference between the three people in the length of time spent watching. Report your results. Be sure to mention which statistical test you used and why (given your results in Q1b), the predictor and outcome variable(s), the appropriate stats reference, and its interpretation in terms of the research question. You should also report and interpret effect size.

[Suggested word count: 80]

Q1d. Regardless of whether your results in Q1c are significant or not, run post-hoc tests. Report on the result, clearly indicating which differences are significant (if any) along with their corresponding p-value and an interpretation of what that means in terms of the research question.

[Suggested word count: 40]

* * *

After some hours, the door finally opens and seven people come in. The first is the bear, who is even bigger than he looked – he is larger anybody LFB has ever seen, and she suspects that is the main person who subdued them. He is followed by the horse (who it turns out is actually a gorgeous unicorn with a rainbow mane), the hippo, an owl, and a cute penguin carrying a snake. The entire group is trailed by what looks, to their astonishment, to be a sentient guitar (that's right, a musical instrument that can walk and talk). *This is a very strange place*, thinks LFB.

Everyone looks at each other awkwardly.

LFB and Foxy are both feeling a little calmer now that it appears nobody is going to try to kill them on sight. Still, the Others seem rather suspicious (not that that is surprising, really).

Everyone introduces themselves: the bear is called Super Size, the unicorn is Rainbow, the hippo is Hugo, the owl is Wol, the penguin is Little Blue, the snake is Sissily, and the guitar is Kevin.

"What are your names?" Super Size asks them.

"LFB," says LFB, trembling.

"What kind of name is LFB?" asks the guitar.

LFB bites her tongue and narrowly avoids asking what kind of guitar is named Kevin, and just says "It stands for Lovable Fluffy Bunny. My mum named me."

The unicorn – Rainbow – shakes her tail and says "It's a lovely name. I like it," and glares at Kevin.

"How about hers?" the snake hisses, pointing at Foxy.

"*She* can answer for herself," Foxy says, bristling a bit. "My name is Foxy. Because I am a fox."

"Okay, okay," says Hugo the hippo. "That's fine. Are you okay? We didn't mean to hurt you when we captured you, we just didn't want to let you get away."

Mollified, Foxy nods. "Head hurts a bit but I'm okay."

"What are you doing here?" the giant owl interrupts.

Trading back and forth, LFB and Foxy tell everyone the whole story -- how they fear they are running out of food, and they wanted to see if the Others were stealing it (at this point LFB trembled a little bit more, and Foxy gave her a reassuring hug) or were having similar problems. As they get into the story, they can't help but notice that most of their listeners seem stunned. The penguin whispers to the unicorn and the giant bear several times during the explanation. When they stop, there is a long silence.

"How do we know you're telling the truth?" the snake, Sissily, finally asks.

"I... don't know," LFB says. "We are, I swear."

After a pause, Little Blue (the penguin) raises her wing. "I have an idea," she says. "We can separate them and then ask them each a bunch of questions. If the answers agree with each other, that's an indication that they aren't lying."

"What kind of questions?" asks the unicorn curiously.

"All sorts," says Little Blue. "A lot of them would be details about their story which they haven't already shared with us – like small details about who was with them or what time they set out or whatever – which if they are telling the truth they would agree on."

Oh dear, thinks Foxy. *This sounds very stressful.*

"I can take the test," LFB volunteers. "I mean, I know that we're telling the truth and I'd like to demonstrate that."

Foxy pauses, and then says, "I can take it as well."

Kevin bursts out suspiciously, "How do we know they won't cheat?"

LFB bristles, but Rainbow the unicorn just smiles. "Don't worry," she says. "We'll separate them right now so they can't coordinate in advance. And we'll have them take it at the same time, in separate rooms, so there is no way to overhear each others' answers."

After a moment, Kevin nods, mollified, and the rest of the Others join in. It seems to be settled.

I hope this works, Foxy thinks nervously. *I know I'm telling the truth, but what if I mess this up? What if their test isn't very good after all?*

But there is no choice – she can't think of anything else that would persuade them better, so she nods and tries to look confident. She can't help but notice that LFB and Little Blue look nervous also, but everyone gets up and the two of them go to separate rooms to take the test.

The tibble d1, which has been loaded for you in the R Markdown document, contains the results of the test. It is described in Table 2 below.

Variable	Description	Table 2: variable description for tibble d1
question	The question number. There are 40 questions total (thus 40 rows). Each question is of the form "On a scale of 0 to 100, how much would you agree with the following statement." (So, for instance, both LFB and Foxy would give a high number for a statement like "we are running out of food in Bunnyland" and a low number for a statement like "we are here in Otherland to hurt you").	
lfb	LFB's answer to that question, on a scale of 0 to 100 (0 indicates complete disagreement, 100 indicates complete agreement)	
foxy	Foxy's answer to that question, on a scale of 0 to 100 (0 indicates complete disagreement, 100 indicates complete agreement)	

Q2 [16% of total mark]

Q2a. The Others are very interested in knowing if there is a correlation between LFB's and Foxy's answers. Based on the data in d1, what would you tell them? In your answer, explain which statistical test you used and why, include the appropriate stats reference, and interpret what this result means about the nature and/or presence of a relationship between these two variables.

[Suggested word count: 40]

Q2b. Instead of using a correlation to answer the question in Q2a, use a linear regression with LFB's answers as the outcome measure and Foxy's answers as the predictor. You do not need to report it; instead, use it to answer the following questions. (i) You should see that the degrees of freedom is the same for the correlation and the linear regression; what other two parameters are also identical? (ii) What parameter of the linear regression is the correlation coefficient *r* related to, and how is it calculated? (iii) What two additional pieces of information does the linear regression give us about the answer to our research question that were not evident from the correlation?

[Suggested word count: 120]

Q2c. Calculate and report the standardised coefficient for the linear regression in part Q2b.

Q2d. The code chunk calculates two new variables called *lfbZnorm* and *foxyZnorm*. Based on the code, explain what it is calculating and what the purpose is. Rerun the analysis in Q2a but use these new variables instead. Is the answer the same or different? Why or why not?

[Suggested word count: 100]

Q2e. Re-run the linear regression from Q2b on the new variables created in Q2d. The slope and intercept parameters are different from the answer in Q2b. Report them, explaining how and why they are different. (You don't need to report the entire regression, just discuss those parameters).

[Suggested word count: 80]

* * *

The Others confer a bit and realise that regardless of the results of the test, over the course of working and talking with LFB and Foxy they have realised that the two are at least reasonably trustworthy. Following a long, whispered conference amongst each other, Rainbow the unicorn steps forward and unties them.

"Sorry for our suspicion. We have been having food problems ourselves," she confides quietly. "We haven't known what to do about it, and are pretty worried."

"Maybe we could help?" LFB offers. "I mean, I don't know much, but perhaps if we compare problems we'll be able to figure out what's going on. We can tell you what we know about our situation too."

Foxy nods and they all share the survey data we saw in previous weeks. The Others share their food data that you went over in the tutorials, and everyone agrees that there is a problem.

"The thing is," Super Size observes (everyone is now very companionable and speaking frankly), "I fear that this is having a lot of bad indirect effects on everything else. People are more irritable and fighting more, they're sick more often, and things like that."

"Do you have any data about that?" LFB asks, curious.

Little Blue thinks, and then nods finally. "We do have a dataset for crime statistics – specifically thefts – from this year and last year. It's our New Year in a few days so that gives us two complete years, pretty much."

Sissily frowns. "I don't understand what data about crimes like stealing would show us."

"Let's see," Little Blue says. Everyone pauses while she shows the dataset to everyone.

"Hmmm," says Foxy, examining it. "It's counting how many thefts of different things – like food or vehicles or whatever – there were in each year. I suppose if food shortages are really a problem, you'd expect to see more food thefts than other kinds of stealing."

Kevin leans over to have a look. "What does the category 'other' mean?" he asks. "My friend Kevin Clark is a string and got kidnapped a few months ago as a prank. Would that count as 'other'?"

"Wait, there's a sentient string in Otherland?" LFB asks incredulously.

Kevin looks up, miffed. "Yes. He's my best friend," he says. "What, do you think a string can't be intelligent? Or a guitar, for that matter?"

"No, no, just curious," LFB backpedals hastily. "All good."

Rainbow whispers to her, "We don't understand it either. Just go with it."

Super Size clears his enormous throat. "I don't think we need to worry too much about the interpretation of what 'other' means," he says tactfully. "The main question is about the food. As long as this dataset is representative then it might be useful."

"It certainly is," Little Blue assures everyone.

The data is in the tibble called dc, which has been loaded for you. It is described in Table 3 below.

Variable	Description	Table 3: variable description for tibble dc
day	Each row corresponds to one theft on one day. Days are coded as *dayN* where *N* indicates the number of the day (e.g., *day121* would mean that day 121 of that year was when the theft in question occurred). Many days do not show up in the dataset; that is because there were no thefts recorded on that day in that year. Some days show up multiple times, if multiple thefts occurred on that day.	
year	This indicates which year the theft occurred on: the *current* one or the *previous* one.	
crime	This indicates the kind of theft that occurred. There are four possibilities: *food* (meaning a food item of some sort was stolen); *tech* (meaning electronics or other tech were stolen, e.g., computers or radios or whatever); *vehicles* (meaning some kind of vehicle was stolen, like a car or bike); and *other* (which is anything else, possibly including sentient string but also including things like books or money).	

Q3 [16% of total mark]

"I made a figure showing the data," Little Blue says helpfully. It is shown below.
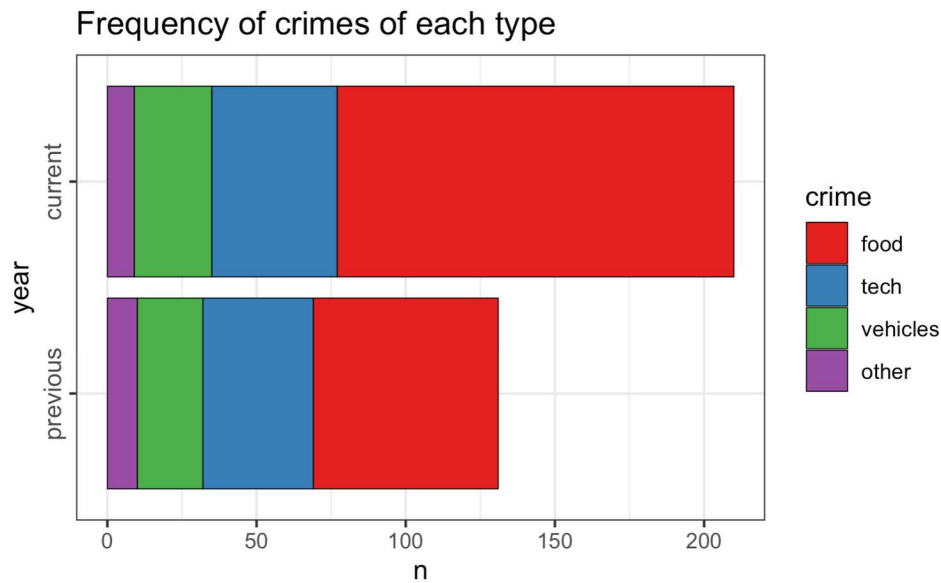
Q3a. Using the tibble called `dcPlotTable` provided in the code chunk, make the figure above. Hint: the geom used is one you have already been taught, but with different arguments; part of the purpose of this question is to evaluate your ability to figure this kind of thing out by using the help files and google. For full credit, your figure should have all the components with everything in approximately the same places. The colours have no transparency. Use the `Set1` palette.

* * *

"Let's ignore the previous year for now," LFB suggests. "In the current year, it sure *looks* like there are more thefts of food than the other categories, but is that significant?"

Little Blue nods. "Good question. Another way of asking it would be whether this is the kind of data we'd expect to see if each kind of theft were equally probable."

Q3b. Considering the current year only, perform the analysis that would answer LFB and Little Blue's question. Be sure to explain what test you did on what variables, include the stats reference, and interpret the result in terms of the research question. You don't need to present the descriptive statistics since they are in the figure. Don't worry about effect size either.

[Suggested word count: 60]

* * *

"Just looking at the current year doesn't tell us what we want to know," Kevin interjects. "Suppose people have always stolen food more, regardless of whether it's running out or not. Like maybe it's just easier to steal or there's a lot more of it or something."

Rainbow glances over with a glimmer of respect in her eyes. "That's a good point, you know."

"You don't need to sound so surprised," Kevin grumbles, but there is a grin in his voice. LFB can't really read his expression – if guitars can be said to have expressions.

"So maybe what we should do is compare the distribution of crimes in the current year to the distribution in the previous year. If they are different, that suggests that something has changed."

Q3c. Compare the thefts from the current year to the previous year using the appropriate statistical test. Be sure to explain what test you did and why, include the stats reference, and interpret the result in terms of the research question. As before, you don't need to present the descriptive statistics since they are in the figure. Don't worry about effect size either.

[Suggested word count: 50]

Q3d. Use the code chunk to display the adjusted residuals from the test in Q3c. What do they suggest about what crime types, if any, are significantly different between the current and the previous year? In your answer be sure to explain why they suggest that.

[Suggested word count: 70]

* * *

"You know what?" Hugo says suddenly. "I just remembered another dataset we could use."

"What is it?"

"The food safety association keeps track of a bunch of information about food over time, and that includes the average price different foods sell for. Maybe we could compare the cost of food this year to the cost of the same food last year."

Hugo's data is in the tibble called `df`, which has been loaded for you. It is described in Table 4.

Variable	Description	**Table 4:** variable description for tibble `df`
names	Each row corresponds to one food item (e.g., *beans, chocolate, eggs, flour*, etc). There are 80 different items total, thus 80 rows. The *names* column indicates the name of the food item for that row.	
previous	This indicates the average price of one unit of that food item in the previous year. For instance, the value of $10.84 for sugar means that one bag of sugar cost, on average, $10.84 last year. Different food items have different units so price comparison across foods isn't that meaningful but the prices can be compared within the same food from year to year because the unit is the same each year.	
current	This indicates the average price of one unit of that food item in the current year. The unit is defined in the same way as for that food in the previous year. Thus, if the current price is higher than the previous price that means the price has gone up and if it is lower that means it has gone down.	

"Wow, that's really cool," says Super Size as he looks at the dataset. "These are the same foods for this year and last year?"

"Yep," says Hugo. "The prices fluctuate but maybe on average they've gone up or down."

"What would that indicate if they did?" asks Foxy.

"Supply and demand, I guess," Rainbow answers. "If the prices went up a lot this year, maybe it's because there's less supply due to less food being available. If they went down, maybe it indicates some other kind of market shift."

"Price changes could occur due to all sorts of things though, not just food shortages," argues Sissily. "It could be inflation or all the grocery stores forming an evil cabal or –"

"True," Hugo interrupts. "We can't draw conclusions about cause based on observational data like this. But it's still useful to see whether the price is changing at all. If it's *not*, that makes it less likely that there are food shortages."

"Okay, I just made a figure showing the data," Little Blue interjects. "Let's see."

"That was fast," LFB says with respect, and Little Blue smiles happily. Everyone clusters around to see the figure. It is shown below, and the code for creating it was given to you in the code chunk.

"Hmm," says Super Size. "Looks like the price has gone up. But is that a significant difference?"

Q4 [9% of total mark]

Q4a. There is one main assumption you need to test before implementing the analysis that would answer Super Size's question. Test that assumption and report the results. Be sure to include the assumption, the name of the test you did, the variable(s) involved, the results including stats reference, whether the assumption was violated, and how you could tell.

[Suggested word count: 40]

Q4b. Given the answer in Q4a, use the appropriate statistical test to implement the analysis that would answer Super Size's question. Report on the results. Don't worry about descriptive statistics, but include the predictor and outcome variable(s), the effect size and its interpretation, and the results including stats reference. You should also explain which statistical test you used and why as well as interpret the result in terms of the research question.

[Suggested word count: 120]

* * *

"That's all fine," Kevin says, "but isn't this dancing around the real issue? The thing that matters is people's health. Who cares if there is less food as long as there's enough for everyone to be healthy, right? So what we need is data about health and how it varies based on different factors."

Little Blue is clearly annoyed. "You keep finding problems with every dataset anyone has," she says. "It's way easier to criticize than it is to actually provide something useful."

"Stop being defensive," Sissily snaps. "It's a good point."

Hugo rolls his eyes. "Both of you, ugh."

Foxy and LFB exchange a look. *It seems the Others are just like normal people*, LFB thinks.

Super Size raises his hands. "It *is* a good point but Little Blue is right, what we need most is all kinds of different data," he says firmly. Foxy is really starting to appreciate his calmness and diplomacy. "Do we have any other datasets we could look at as well?"

The owl, who has been mostly silent so far, raises a wing. "I know the health department recently ran a survey with a representative sample," she says. "It's not huge, but they did ask about health, along with other things like income and exercise level. Maybe that might be useful?"

"It doesn't have anything about diet or nutrition," Hugo says. "But it's better than nothing." The data is in the tibble called dh, which has been loaded for you. It is described in Table 5.

Variable	Description
	Table 5: variable description for tibble dh
person	Each of the 120 rows corresponds to one person, labeled *personX*. The *X* is an arbitrary number between 1 and 120 that was randomly assigned so each person had a unique identifier in the dataset.
health	The health rating of that person on a scale from 0 to 100. It is a comprehensive health rating that covers many different aspects of health. All we need to know is that lower rating means poorer health: 0 means "almost dead" and 100 means perfect health.
income	The annual income of that person in thousands of dollars. Thus, 55.5 means they make $55,500 a year.
inCat	A grouping variable based on income. A person is put in the income category *poor* if their income is less than the median and in the category *rich* if their income is greater than the median.
exercise	The average amount of exercise each person gets in terms of average calories (kCal) burnt through exercise per day. Thus, 200 would mean they burn 200 calories.
exCat	A grouping variable based on exercise. A person is put in the exercise category *low* if they burn fewer calories on average each day than the median and *high* if they burn more than the median.

"Wow, there's a lot of data," Rainbow observes. "Where do we start?"

"I have an idea," says Super Size. "Let's see if average health varies between the two income categories (*poor* vs *rich)* or the two exercise categories (*low* vs *high*) or if there's any interaction."

Q5 [12% of total mark]

Q5a. Create a figure that you can use to display the information Super Size wants to see. Part of the purpose of this question is for you to determine about what kind of figure is best given this specific aim and these specific variables, so you must decide which geom(s) to use, what variables to include, and how to design everything. Your mark will be based on whether your figure displays the information Super Size wants to observe in a way that is both clear and informative.

Q5b. Run the appropriate statistical test to evaluate whether income category or exercise category are significant predictors of health rating, including whether there is an interaction. You do not need to test any assumptions and can act as if none have been violated. In your report, don't worry about including descriptive statistics but do include an explanation of which statistical test you used, the stats reference(s), the predictor and outcome variable(s), and the interpretation of this data in terms of the research question. Don't worry about effect size or post-hoc tests.

[Suggested word count: 140]

* * *

"You know," LFB says, "if we want to look at how income and exercise are related to health, we should use the predictor variables *income* and *exercise* instead of the income and exercise categories. Those have the exact values instead of only two groups each."

"I don't see why that would make a difference," Rainbow says doubtfully.

"I think she's right," Little Blue says. "That's a much better idea."

Q6 [19% of total mark]

Q6a. Do as LFB suggests: run the appropriate statistical test to evaluate whether exercise and/or income are significant predictors of health rating. Don't include an interaction term. You do not need to test any assumptions and can act as if none have been violated. In your report, don't worry about including descriptive statistics but do include an explanation of which statistical test you used, the stats reference(s), the predictor and outcome variable(s), the effect size of the model as a whole, and the interpretation of this data in terms of the research question. Include a description and interpretation of any significant unstandardised coefficients but don't worry about standardised ones or the intercept.

[Suggested word count: 140]

Q6b. Run the same analysis as in Q6a but this time include an interaction term. (i) Report the results of the interaction term only, including its stats reference as well as interpretation in terms of the research question. (ii) Compare the effect size of this model to the one in Q6a. What does this mean about which model, if any, fits the data better? (iii) Describe how the unstandardized coefficients for each of the two predictors have changed, and explain why. Include an interpretation of these predictors in terms of the research question.

[Suggested word count: 140]

Q6c. Does your analysis in Q5b and Q6b lead to the same qualitative conclusion about which factors are significantly associated with health rating? If so, explain on what basis one should decide which analysis to report. If not, explain why the two are different.

[Suggested word count: 100]

Q7 [6% of total marks]

Q7a. For each of (i), (ii), and (iii) in the table below, indicate which of the two statistical references (A or B) is *not* possible. Use your knowledge about each test statistic and how it is related to degrees of freedom and/or p-values to answer this. You don't need to do any coding or calculations.

	Option A	Option B
Q7a(i)	$\chi^2(2) = 0,\ p = 1$	$\chi^2(2) = -8.69,\ p = .013$
Q7a(ii)	$t(161.2) = 5.06,\ p < .001$	$t(166) = 86.54,\ p = .951$
Q7a(iii)	$F(2,132) = 0.85,\ p < .001$	$F(2,247) = 3.78,\ p = .024$

Q7b. For each answer in Q7a, explain it by describing intuitively what the test statistic captures and thus why you chose the answer you did. Be sure to make reference to the degrees of freedom, test statistic, and/or p-value as appropriate to justify your answer.

[Suggested word count: 190]

Q8 [2% of total mark]

These marks are free as long as you say anything! Who is your favourite character in the Bunnyland story, and why? (This can be anyone, whether from Bunnyland or Otherland or whatever. There is no word limit here, say as much or as little as you want).