STATS 369 test

There are 6 questions.

Answer all the questions

There are 64 marks in total

You have 75 minutes

Upload your answers to Canvas as a pdf, ensuring that it is clear what question you are answering.

Integrity Statement

By completing this assessment, I agree to the following declaration.
I understand the University expects all students to complete coursework with integrity and honesty. I promise to complete all online assessment with the same academic integrity standards and values. Any identified form of poor academic practice or academic misconduct will be followed up and may result in disciplinary action.
As a member of the University's student body, I will complete this assessment in a fair, honest, responsible and trustworthy manner. This means that:

- I declare that this assessment is my own work.
- I will not seek out any unauthorised help in completing this assessment.
- I am aware the University of Auckland may use plagiarism detection tools to check my content.
- I will not discuss the content of the assessment with anyone else in any form, including Canvas, Piazza, Facebook, Twitter or any other social media or online platform within the assessment period.
- I will not reproduce the content of this assessment anywhere in any form at any time.
- I declare that I generated the calculations and data in this assessment independently, using only the tools and resources defined for use in this assessment.
- I will not share or distribute any tools or resources I developed for completing this assessment.

1. Answer the questions below in the context of this course (9 marks total, 3 marks each)

(a) Explain when cross-validation is preferable to using a test-train spilt and when a test-train split is preferable to cross-validation.

(b) What is mean-squared prediction error and why is it hard to estimate?

(c) Define apparent error (defining any terms used) and discuss whether it is a useful estimate for MSPE.

2. Consider the sales data frame, for which the header is given, and the code fragment below. Assume sales has at least one entry for every month from 2015 until now.

```
## # A tibble: 9004 x 6
##   date   item_name   item_cost  item_code   discount  notes
##   <chr>  <chr>       <dbl>      <dbl>       <dbl>.    <chr>
## # … with 9004 more rows
```

```
sales %>%
separate(date, into=c("year","month",“day”,“other”), sep = “-“) %>%
filter(year == “2019”)
group_by(month) %>%
summarise(inc = sum(item_cost), trans = n(), disc = mean(discount))
```

What columns will the output of this code have, what data will be in each column, and how many rows will it have? (7 marks)

3. In a regression framework, for **each** scenario below, name **two** methods covered in the course so far that are suitable for addressing the scenario. Briefly discuss their pros, cons, and how they compare with each other.
You can name six different methods, or repeat if appropriate.
(15 marks total, 5 marks each)

a) ***p >> n***: The number of predictors *p* is much larger than the number of observations *n*

b) **Non-linearity**: The relationship between predictors and response is non-linear. (5 marks)

c) **Heteroscedasticity**: The variance of the error term, $\sigma^2 = \mathrm{Var}(\epsilon)$ is not constant. (5 marks)

4. You are interested in predicting presence or absence of a type of cancer (Y) using protein measurements (X) from blood samples on 200 people. There are about 2000 X variables measured, but the biochemists only give you a subset of 100 of them. You do cross-validated model selection to end up with a predictive model and an estimated error rate from cross-validation.
Would you expect the estimated error to be biased (and why or why not) if:

(a) The 100 variables were chosen because they were correlated with cancer status in the same 200-person data set (3 marks)

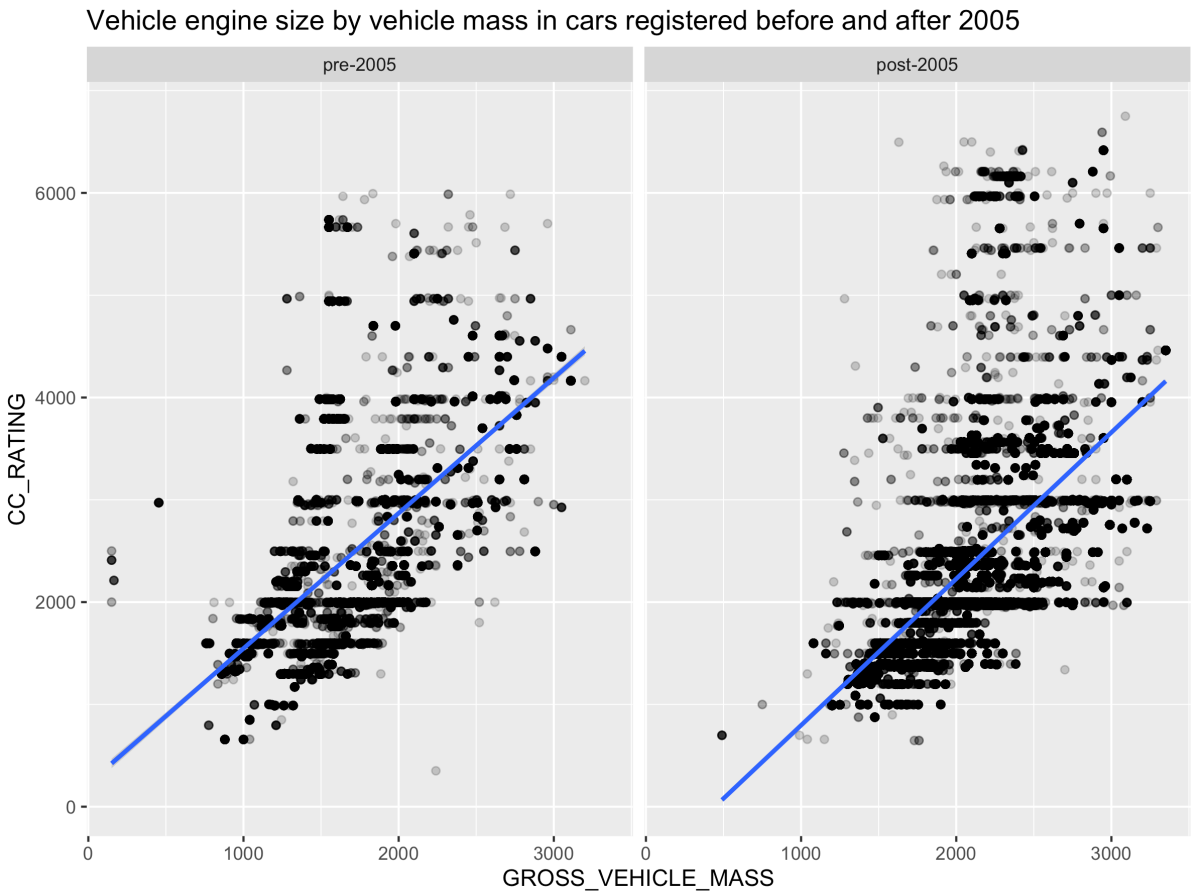
(b) The 100 variables were chosen because they had predicted cancer in previous research (3 marks)

(c) The 100 variables were pre-selected based on a prior cross-validation process on the same 200-person data set (3 marks)

5. The following plot is made from the New Zealand vehicles data that we have seen several times in lectures.

Name the ggplot commands that could have been used to make this plot, including any geoms or aesthetic mappings, and explain what each does.

(10 marks)



6. Download the files q6.Rmd and strains.csv. Answer the questions in the Rmd file and knit the result to html.
Upload both the html and Rmd
(14 marks total)