

QMS210: Applied Statistics for Business SPSS GROUP PROJECT WINTER 2022

DUE: April 7, 2022, by 11:59 pm.

MARKS: Total marks = 100 (or 14% of the final grade)

PENALTY: There will be a 25-mark penalty (or 25% of the project mark) for every day after the due date (including weekends).

Notes:

1. **No handwritten reports will be considered for marking purposes.**
2. **Submit one copy of the project per group with a cover page, report, SPSS (or Excel) output file and indicate your group number and the names and student numbers of all the group members on the cover page. Failure to indicate your group number will result in a zero grade. There will be penalties for the inclusion of unnecessary information.**
3. **Present your solution for Hypothesis testing with the two approaches according to the template shown in class. All relevant SPSS or Excel outputs (and ONLY relevant outputs) must be included with proper labelling of graphs and charts.**
4. **Upload all your SPSS data files or excel files with your report online via D2L under Group Discussion.**
5. **Each group must submit a single file of your report containing the 5 questions in either Word or PDF format and all the SPSS output files in spv format or Excel output files for the 5 questions. Failure to follow this may result in a zero mark.**
6. **Each group has a unique data set. It is your responsibility to use your unique data set. You must inform your professor and ask for a new data set if you change group. Failure to follow this may result in a zero mark.**
7. **Read this entire document carefully.**

You will analyze a data set that comes from a random sample of the Usage-Based Insurance Synthetic Dataset. This data set was provided by the following published article.

So, B.; Boucher, J.-P.; Valdez, E.A. Synthetic Dataset Generation of Driver Telematics. *Risks* **2021**, *9*, 58.
<https://doi.org/10.3390/risks9040058>

Your data set was randomly resampled from the big data set that consists of more records. So, please be aware that each group will have its own data set. You will only focus on the relevant variables and answer each question.

Description of Data Variables:

| Variable | Description |
|--------------------|--|
| Insured.age | Numerical, drivers' age. |
| Insured.sex | Categorical, drivers' gender, male or female. |
| Car.age | Numerical, how many years old for the vehicles. |
| Marital | Categorical, Married or Single |
| Car.use | Categorical, Private use, Commute, or Commercial use. |
| Credit.score | Numerical, Drivers' credit score. |
| Region | Categorical, Urban or Rural. |
| Annual.miles.drive | Numerical, how many miles driven per year. |
| Years.noclaims | Numerical, how many years that the drivers have no claims. |
| Territory | Categorical, which residential area, labelled by a number. |
| Claim.occurrence | How many claims within current policy year. |
| AMT_Claim | If there is a claim, the amount of the total claim reported. |

How the Project is graded

Your submission will be graded based upon the following factors: substance, presentation, accuracy, grammar and clarity. A demonstration of effort is the driving force of this assignment. Assignments will be compared to discern levels of effort and excellence.

As a minimum, your report must include the following:

1. Title page: [1] title [2] submission date [3] group number and the file name of the data set used [4] names of each group member with their Ryerson ID student number and [5] course code and section (e.g., **QMS210 - Section 01**).
2. Your project must be submitted online via D2L under Assessments ⇒ Assignment.
3. The answer to each question will begin on a new page under the statement of the question.
4. Cut and paste all relevant SPSS or Excel outputs in the write-up section at the bottom of your answer to each question. Do not send the reader to appendices to find them.
5. A complete write up of your chosen hypothesis test must include your assumptions, analysis of results and your conclusions. **You must use both approaches (critical value and p-value approaches) to make your statistical decision.**

6. Not using the **exact** data set assigned to your group will result in getting a zero mark for the project. Please be informed that each group has a different data set. If you use other group's data set, both groups will get a zero mark and will be charged with academic misconduct".
7. **All data analyses must be done with SPSS and/or Excel. Only critical values can be found using the recommended calculator or Excel.**
8. **Number the pages of your report.**

Group Size

This project can be done in groups with 3 to 5 members only. **This means that the project report must be a result of team effort.** It is your responsibility to find your group members online via D2L under Communication → Group. Before **Feb 17**, you can self-enroll in a group. After March 7, your instructor (or D2L) will assign you a group number randomly.

Solutions for some potential problems with your group:

1. If any member(s) does/do not participate in the group work, the others in the group have the right to remove them after warning them and also informing the professor. The removed member(s) must form another group of at least 3 and inform the professor to get new data.
2. The dropout will have to join or form another group. The student is not allowed to form a group of less than 3. Failure to form a group may result a zero mark.
3. Any newly-formed groups must obtain a new data set from the professor. Failure to do so will result in a zero mark. It is your responsibility to get a new data set.
4. Any detection of plagiarism in the report will be charged with academic misconduct and all groups will receive a zero mark. You must use Turnitin when you submit your project on D2L.
5. If your name is not in the final report, you must email all the members in your group and the professor with proof of participation.
6. The latest date to form groups is 2 weeks before the submission date. After this date, D2L will not be able to form new groups. (Hence if you are not in a group 2 weeks before submission date, you will receive a zero on the project, since the project must be done in groups of 3-5 only).
7. Refer to the group work ethics guidelines posted on D2L and course outline.

6.

THERE ARE 5 QUESTIONS in this project. The following table shows the naming convention for the data set of each group. Please state the name of the data set on the cover page of your project.

| DATA ASSIGNMENT | |
|-----------------|----------------------|
| Group # | Name of the data set |
| 1 | QMS210Group_1 |
| 2 | QMS210Group_2 |
| 3 | QMS210Group_3 |
| 4 | QMS210Group_4 |
| 5 | QMS210Group_5 |
| 6 | QMS210Group_6 |
| 7 | QMS210Group_7 |
| 8 | QMS210Group_8 |
| 9 | QMS210Group_9 |
| 10 | QMS210Group_10 |
| 11 | QMS210Group_11 |
| 12 | QMS210Group_12 |
| 13 | QMS210Group_13 |
| ...etc | ...etc |

IMPORTANT:

Each GROUP HAS ITS OWN unique Data Set. The data set assigned to your group consists of 500 records of the driver's personal, driving amount, and claim information.

Question 1 (24 marks) (8+8+8=24 marks)

- Based on the assigned data to your group, construct a **pie** chart for the variable “**Car.use**” that represent the distribution of how cars are used. Your pie chart should show the category names with the percentage breakdown that is, data labels in percentage. Include the chart in your report.
- Based on your pie chart, identify the most popular type of use.
- Make a contingency table for the two variables: “**Insured.sex**” and “**Car.use**”. State the most frequent use of the car by gender.
- Draw a scatter plot on “**Insured.age**” and “**Years.noclaims**”. Comment on their relationship based on the scatter plot.

Question 2 (22 marks) (12+6+4=22 marks)

- a) Find the measures of central tendency (mean and median) for the two variables: "**Insured.age**" and "**Credit.score**". Discuss the shape of these two distributions. Which measure of central tendency is the best to represent the "**Credit.score**" data: the mean or the median? (Hint: Use 10% rule.) Discuss your rationale for the choice.
- b) Find the measures of variability (range, IQR, variance and standard deviation) for the two variables: "**Insured.age**" and "**Credit.score**".
- c) Which one of the two variables, "**Insured.age**" or "**Credit.score**", is relatively more variable than the other? (Hint: use the CVs).

Question 3 (18 marks) (10+4+4= 18 marks)

- a) Use the variable "**Annual.miles.drive**" to construct the confidence intervals for the estimate of the population mean (i.e., mean of annual miles driven), by gender, i.e. by male and female drivers, at 96% levels. Interpret your confidence intervals.
- b) Did you make any assumptions when constructing your confidence intervals? If yes, which assumptions; if not, why?
- c) If insurance companies charge premiums based on the annual miles driven, based on your finding, is it warranted for insurance companies to charge male drivers higher premiums based on this data set? Explain.

Question 4 (13 marks) (13 marks)

Consider the claim that the average credit score of a driver in the given population is 780. Use the data collected for the variable "**Credit.score**" to test this claim by both critical value approach and p-value approach. (Use the 5% level of significance).

Question 5 (23 marks) (15+4+4=23marks)

- a) Based on your data set, is the average years of no claims ("**Years.noclaims**") of male drivers is significantly MORE THAN the average years of no claims ("**Years.noclaims**") of female drivers? Test both critical value and p-value approaches at the 8% level of significance. Note that SPSS only performs a 2-sided test.
- b) Provide possible reasons why you should expect to find a significant statistical difference between the gender.
- c) Based on your findings, should the insurance company charge higher premiums based on the gender? Explain your reasoning.

SPSS PROJECT HINTS: avoid these pitfalls

- The most common and biggest error is to assign one question to each person and put all parts together. The outcome is almost always of very poor quality and receives a very low grade. Our exams & tests also assume that each of you is an expert in all facets of this project. You must check each other's work in your group- and fully understand it. It is a TEAM effort.
- The 2nd most common error is to postpone the assignment so late that you do not have time to complete it. That is a sure way to do badly in this course.
- The 3rd most common mistake is to fail to monitor your team members. You must learn to manage teams and make sure that you have all the data and reports at the same time.
- You misread the question.
- You used the wrong test (e.g., Using a Z test instead of a t test).
- Your hypothesis was in the wrong direction (or H_1 has $>$ or $<$ instead of \neq).
- The null hypothesis or the alternative hypothesis (or both) was wrong.
- You came to a wrong conclusion.
- You used the wrong data (or Incorrect inputs).
- A hypothesis with a μ or p or has the wrong one.
- H_a contains one of $\{=, \leq \text{ or } \geq\}$ OR H_0 contains one of $\{>, <, \text{ or } \neq\}$.
- You used sample statistics in your hypotheses.
- Failed to check the requirements to use a test
- Misread p-values or comparison of p to α is wrong.
- Reaching a wrong conclusion, i.e., rejecting H_0 , when $p > \alpha$.
- There is no statistical decision (or a wrong one).
- There is no managerial conclusion (or a bad one).
- The test is a one-sided test (not 2-sided).
- Not taking $\frac{1}{2}$ of the Sig value from SPSS for a 1-sided test.
- You failed to state the problem and/or define the variables.
- A printout of your DATA SET IS MISSING! It had to be included!
- Forget to discuss or check for normality.