

## **QMS230: Statistics for Finance and Accounting R GROUP PROJECT WINTER 2021**

**DUE: April 11, 2021 by 11:59pm.**

**MARKS: Total marks = 25 (or 10% of the final grade)**

**PENALTY: There will be a 10% of the project -mark penalty for every day after the due day (including weekends) that the project is late.**

**Notes:**

1. No handwritten reports will be considered for marking purposes.
2. Submit one copy of the project report per group, R script file and indicate your group number and the names and student numbers of all the group members. Failure to indicate your group number will result in a zero grade. There will be penalties for the inclusion of unnecessary.
3. Present your solution for Hypothesis testing according to the template shown in class. All relevant R outputs must be included.
4. Upload your R script file together with your report online via D2L under Group Discussion.
5. The report must be a single file and in PDF format only. Failure to follow this may result in a zero mark.
6. Read this entire document.

**You will analyze dataset comes from a random sample of the National Collision Database for 2017.** This data was retrieved from the Government of Canada website. It includes all motor vehicle collisions in Canada on public roads in 2017 which have been reported to the police. This data set contains a variety of data variables which are summarized in the data dictionary posted on the Government of Canada website.

Source:

Government of Canada. (2017). *National Collision Database*.

Your data set was randomly resampled from a big data set that consist 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:**

<b>Variable</b>	<b>Description</b>
Year	
Month	01-January up until 12-December
Day of week	01-Monday up until 07-Sunday
Collision hour	00-Midnight to 0:59 to 23-23:00 to 23:59
Collision severity	1 – fatal or 2 – non-fatal
Number of vehicles involved in collision	
Weather condition	1 – clear and sunny, 2 – cloudy, 3 – raining, 4 – snowing, 5 – freezing rain, 6 – limited visibility, 7 – strong wind
Road surface	1 – dry, 2 – wet, 3 – snow, 4 – slush, 5 – icy, 6 – sand/gravel/dirt, 7 – muddy, 8 – oil, 9 – flood
Road alignment	1 – straight and level, 2 – straight with gradient, 3 – curved and level, 4 – curved with gradient, 5 – top of hill, 6 – bottom of hill
Vehicle ID	
Vehicle type	<b>See Below</b>
Vehicle model year	
Person ID	
Person sex	01 – Female, 02 – Male
Person age	
Person position	<b>See Below</b>
Medical treatment required	1 – no injury, 2 – injury, 3 – fatality
Safety device used	01 – none, 02 – safety device used, 09 – helmet worn, 10 – reflective clothing worn, 11 – helmet and reflective clothing, 12 – other safety device used, 13 – no safety device equipped
Road user class	1 – Motor Vehicle Driver, 2 – Motor Vehicle Passenger, 3 – Pedestrian, 4 – Bicyclist, 5 - Motorcyclist

**Vehicle Type**

<b>Code</b>	<b>Description</b>
01	Light Duty Vehicle (Passenger car, Passenger van, Light utility vehicles and

	light duty pick up trucks)	
05	Panel/cargo van <= 4536 KG GVWR	Panel or window type of van designed primarily for carrying goods.
06	Other trucks and vans <= 4536 KG GVWR	Unspecified, or any other types of LTVs that do not fit into the above categories(e.g.. delivery or service vehicles, chip wagons, small tow trucks etc.)
07	Unit trucks > 4536 KG GVWR	All heavy unit trucks, with or without a trailer
08	Road tractor	With or without a semi-trailer
09	School bus	Standard large type
10	Smaller school bus	Smaller type, seats < 25 passengers
11	Urban and Intercity Bus	
14	Motorcycle and moped	Motorcycle and limited-speed motorcycle
16	Off road vehicles	Off road motorcycles (e.g. dirt bikes) and all terrain vehicles
17	Bicycle	
18	Purpose-built motorhome	Exclude pickup campers
19	Farm equipment	
20	Construction equipment	
21	Fire engine	
22	Snowmobile	
23	Street car	

#### Person Position

Code	Description
11	Driver
12	Front row, center
13	Front row, right outboard, including motorcycle passenger in sidecar
21	Second row, left outboard, including motorcycle passenger
22	Second row, center
23	Second row, right outboard
31	Third row, left outboard
32	Third row, center
33	Third row, right outboard
etc.	

96	Position unknown, but the person was definitely an occupant	
97	Sitting on someone's lap	
98	Outside passenger compartment	e.g. riding in the back of a pick-up truck
99	Pedestrian	

### 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 plus their student number, [5] course code (i.e.: **QMS230**) [6] Submitted to "Instructor's name"
2. Your project must be submitted online via D2L under Assessments==Assignment.
3. The answer to each question will begin on a new page. State the question (cut and paste).
4. Cut and paste all relevant R 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 will use p-value to make your statistical decisions.**
6. Not using the **exact** dataset assigned to your group will result in getting a zero mark for the project. If you use data from another group, both your group and the other group will receive a zero mark. The data for each group is for their group's use only.
7. **All data analyses must be done with R. Only the critical values can be found using the recommended calculator or R.**

### Group Size

This project can be done in a group. **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 Discussion Board. Your instructor (or D2L) has already assigned you a group number.

**THERE ARE FIVE QUESTIONS in this project. The following table shows the naming convention for the data set for each group. Please state the dataset name on your project.**

DATA ASSIGNMENT	
Group #	Dataset
1	QMS230Group_1
2	QMS230Group_2
3	QMS230Group_3
4	QMS230Group_4
5	QMS230Group_5
6	QMS230Group_6
7	QMS230Group_7
8	QMS230Group_8
9	QMS230Group_9
10	QMS230Group_10
11	QMS230Group_11
12	QMS230Group_12
13	QMS230Group_13

**IMPORTANT:**

Each GROUP HAS ITS OWN unique Data Set. Your group will be assigned a set of data that consists of 600 records of the motor vehicle collisions. Individual projects (teams of 1) are NOT permitted.

If you encounter a problem in using R to get certain output, you can google it. This project also aims at building problem solving skills.

**Question 1 (5 marks)**

a) Based on the data provided, construct a **pie** chart for the variable “**Road user class**” that involved in collision. Your pie chart should show the **category names with the percentage breakdown**, that is, data labels in percentage. Include the chart in your report.

b) Based on your pie chart, identify the most popular “**Road user class**” that involved in collision.

c) Make a two ways contingency table for the two variables: “Collision severity” and “Weather condition”, to summarize the joint frequency values. State the most severe collision by weather.

**Question 2 (5 marks)**

- a) Using R functions, find the measures of central tendency (mean and median) for the two variables:  
"Person Age" and "collision Hour". Discuss the shape of these two distributions by plotting their histogram (you can use the default setting of the class width for the histograms.). Based on the shape of distributions, which measure of central tendency is best to represent the "Person Age" data: the mean or the median? Discuss your rationale for the choice.
- b) Using R functions, find the measures of variability (range, IQR, sample variance and sample standard deviation) for the two variables: "Person Age" and "Collision Hour".
- c) Which of the two variables, "Person Age" or "Collision Hour" is relatively more variable than the other? (Hint: use the CVs)
- d) Using the central tendency measures to describe these two variables, and to state at what time and at what age, a collision is more likely to happen.

**Question 3 (5 marks)**

- a) Use the variable "**Person age**" to construct the confidence intervals for the estimate of population mean "**Person age**" involved in a collision on a **cloudy day** (Code=2), at both 95% and 99% levels. You must compute Confidence Interval using R. Interpret your confidence intervals.
- b) Did you make any assumptions when constructing your confidence intervals? If yes, which assumptions; if not, why?
- c) Consider the claim that the average collision hours of collision at the time the data was collected was equal to 15 hours. Use the variable "Collision Hours" to test this claim. You must conduct the hypothesis test using R. (Use the 5% level of significance).

**Question 4 (5 marks)**

- a) Based on your data, is the average age of the driver who involved in a fatal accident is significantly **MORE THAN** the average age of driver who involved in a non-fatal accident? Test at the 10% level of significance, using R to conduct the test.
- b) Provide possible reasons why you should expect to find a significant statistical difference between the ages of these 2 groups.
- c) Based on your findings, should the insurance company charge more for the young driver? Explain your reasoning.

**Question 5 (5 marks)**

By looking at the variables in this national collision database, what other statistical analysis can be done in order to answer some specific questions you have in mind? Make a detailed proposal of statistical analysis and justify your thought. State a plan on how you are going to conduct such statistical analysis using the methods and techniques that you learned from this course. Write a paragraph with approximately, 500 words.