Hazardous Driving Areas in Canada

By Ruei-Hung Chen, Xuechun Qian, Zhenyu Xuan and Ziyang Li

Tutorial Section: *0101E*
Group Number: *E1*

Winter 2018

Introduction

Definition: The provinces in Canada which have SeverityScores greater than the mean of SeverityScore are considered to be hazardous area.

Based on our definition: we believe the province with the greatest median of SeverityScore is the most hazardous

Accident distribution in Canada:

Methods: qmplot()

Methods:

1.) To understand the data:

- Examine a new dataframe that only contains the provinces in Canada.
- Analyze a new dataframe that only contains the provinces with the SeverityScore greater than the mean of the SeverityScore among Canada.

2.) Hypothesis Test

- Compare the differences between provinces.

3.) Sampling Distribution

- Simulate possible values (5000 times) of the test statistic.

4.) Regression Model

- Evaluate the relationship within SeverityScore, trucks and percentage of heavy duty trucks.

5.) Outside data

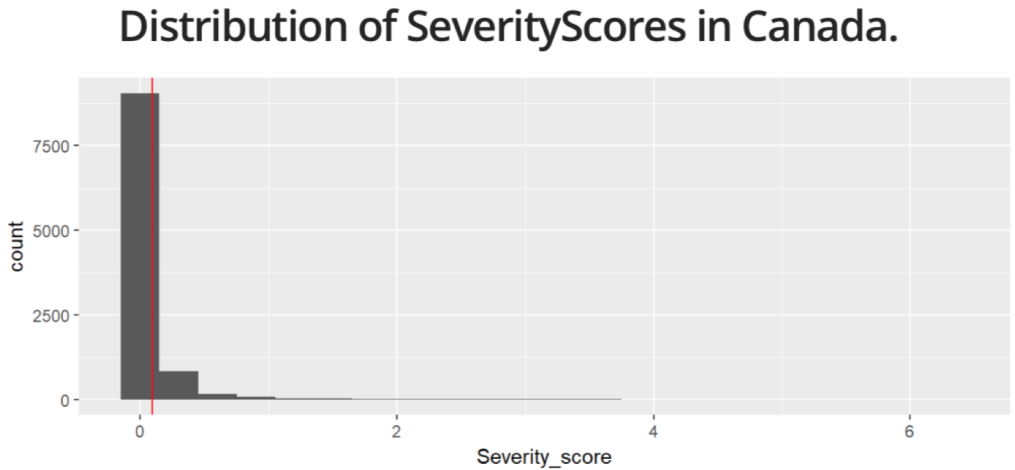
- Assess the percentage of heavy duty trucks among all registered vehicles.

Citation:
1) geotab: Hazardous Driving Area. (n.d.). Retrieved March 31, 2018, from https://data.geotab.com/urban-infrastructure/hazardous-driving
2)STat canada 1: (2017, June 29). Retrieved March 31, 2018,fromhttp://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=4050004&&pattern=&stByVal=1&p1=1&p2=37&tabMode=dataTable&csid=
3)MOOSE:MooseStatistics.(n.d.).RetrievedMarch31, 2018, from http://sopacnl.com/statistics/

Defining Hazardous Driving

SeverityScore Range:

```
## [1] 0.0005 6.3259
```

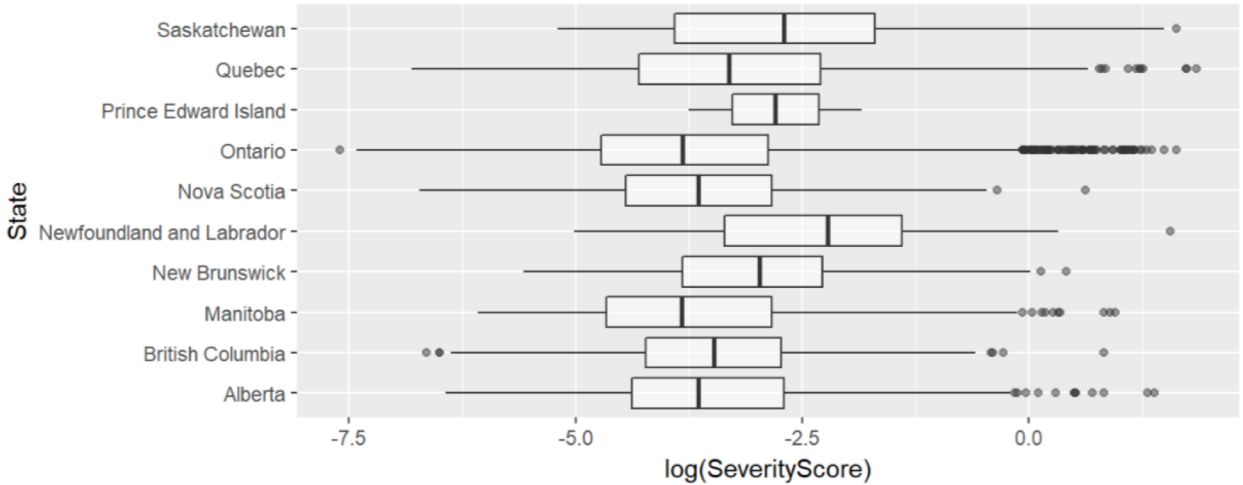


- *Red line*: The mean of SeverityScore *0.0912*.

The Boxplot Distribution of SeverityScore

- ➔ Numerical variable: SeverityScore
- ➔ Categorical variable: State (provinces in Canada)

- Filter() only the accidents with SeverityScore greater than the mean of SeverityScore.
- Use Log() and coord_flip() to transform data.



Summary of median, mean, maximum, minimum

State	median	mean	max	min	N
Alberta	0.02610	0.1085694	3.9876	0.0016	477
British Columbia	0.03105	0.0659823	2.2843	0.0013	666
Manitoba	0.02160	0.0863245	2.5803	0.0023	601
New Brunswick	0.05105	0.1301392	1.5010	0.0038	120
Newfoundland and Labrador	0.10820	0.2724061	4.7400	0.0066	49
Nova Scotia	0.02600	0.0602934	1.8583	0.0012	243
Ontario	0.02190	0.0808008	5.0800	0.0005	5909
Prince Edward Island	0.09065	0.0906500	0.1578	0.0235	2
Quebec	0.03635	0.1152767	6.3259	0.0011	2128
Saskatchewan	0.06680	0.4391435	5.0761	0.0055	46

- **Newfoundland and labrador** has the highest median of 0.108.
- **Saskatchewan** has the highest mean of 0.439.

We target the most hazardous province based on median, because means are susceptible to extreme values.

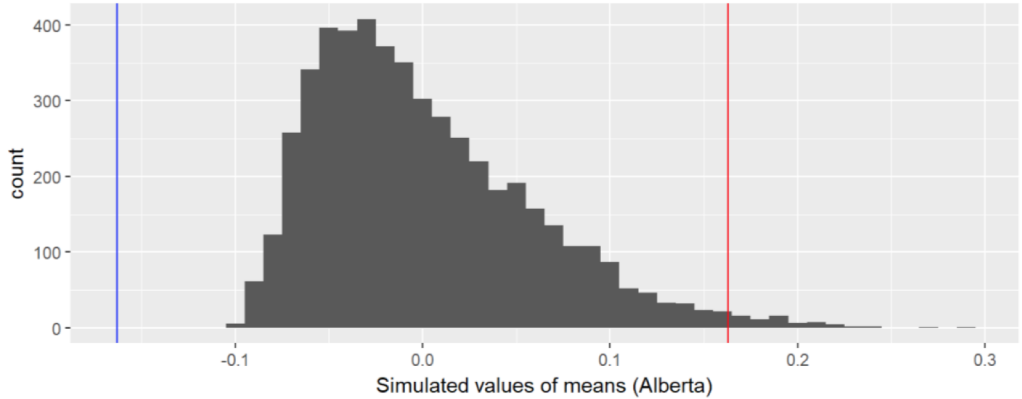
However, our sampling distribution is based on mean, because each province has different sample size, median is not suitable.

Sampling Distribution 1:

Newfoundland Vs. Alberta

$$H_0 : \mu_{nfl} = \mu_{alberta} \qquad H_a : \mu_{nfl} \neq \mu_{alberta}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{alberta} = 0.272 - 0.109 = 0.163$$



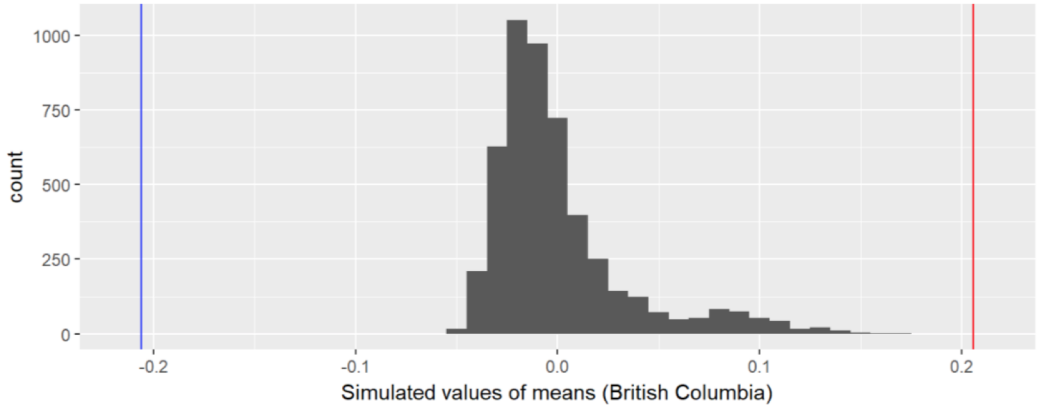
- *P-value = 0.0144*

Sampling Distribution 2:

Newfoundland Vs. British Columbia

$$H_0 : \mu_{nfl} = \mu_{bc} \qquad H_a : \mu_{nfl} \neq \mu_{bc}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{bc} = 0.272 - 0.066 = 0.206$$



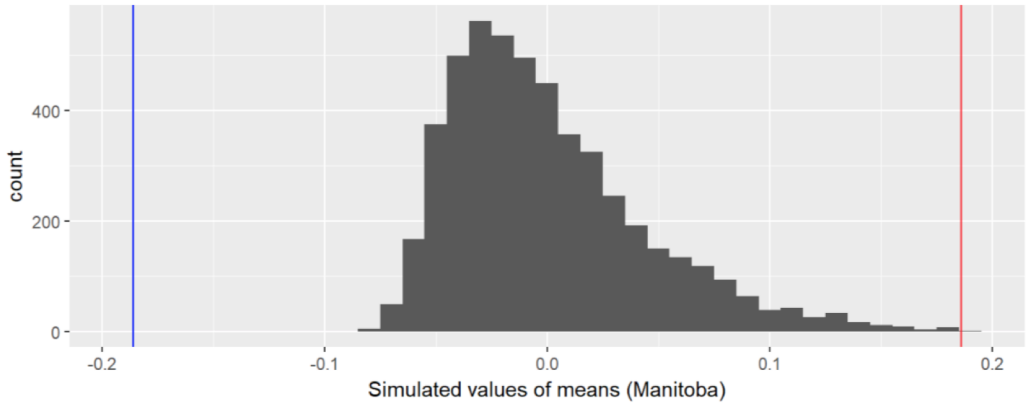
- *P-value = 0.0*

Sampling Distribution 3:

Newfoundland Vs. Manitoba

$$H_0 : \mu_{nfl} = \mu_{mb} \qquad H_a : \mu_{nfl} \neq \mu_{mb}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{mb} = 0.272 - 0.086 = 0.186$$



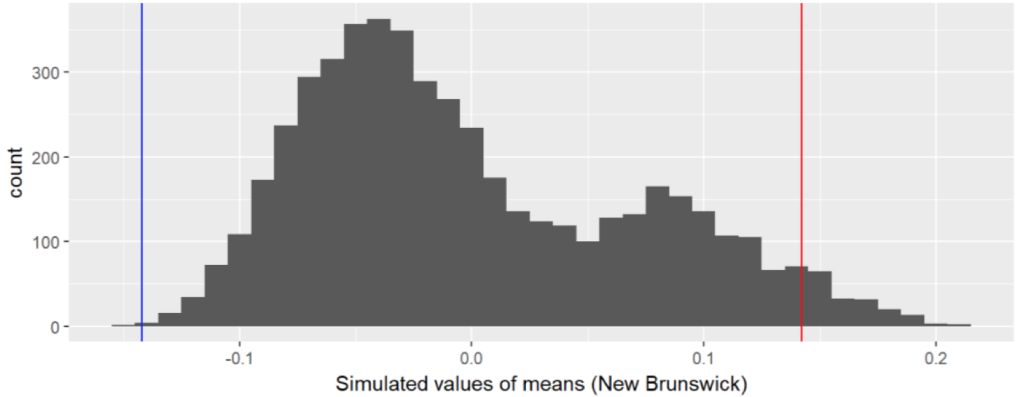
- *P-value = 0.0*

Sampling Distribution 4:

Newfoundland Vs. New Brunswick

$$H_0 : \mu_{nfl} = \mu_{nb} \qquad H_a : \mu_{nfl} \neq \mu_{nb}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{nb} = 0.272 - 0.130 = 0.142$$



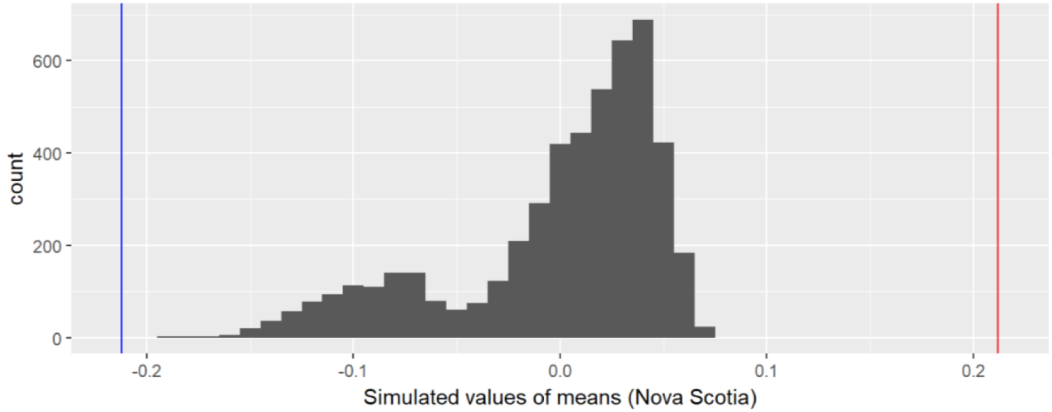
- *P-value = 0.0347*

Sampling Distribution 5:

Newfoundland Vs. Nova Scotia

$$H_0 : \mu_{nfl} = \mu_{ns} \qquad H_a : \mu_{nfl} \neq \mu_{ns}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{ns} = 0.272 - 0.060 = 0.212$$



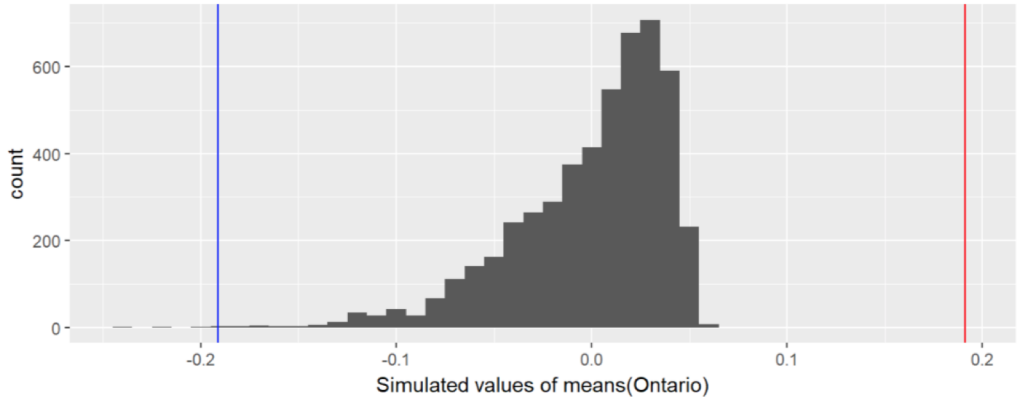
- *P-value = 0.0*

Sampling Distribution 6:

Newfoundland Vs. Ontario

$$H_0 : \mu_{nfl} = \mu_{on} \qquad H_a : \mu_{nfl} \neq \mu_{on}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{on} = 0.272 - 0.081 = 0.191$$



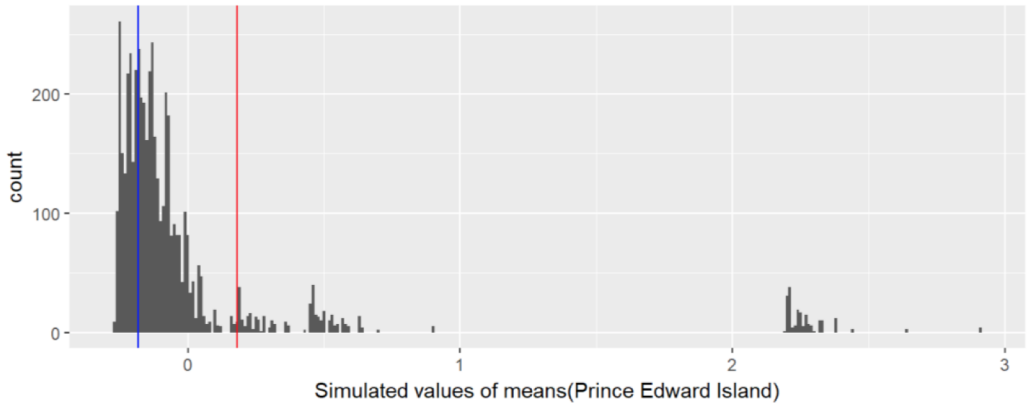
- *P-value = 6e-04*

Sampling Distribution 7:

Newfoundland Vs. Prince Edward Island

$$H_0 : \mu_{nfl} = \mu_{pei} \qquad H_a : \mu_{nfl} \neq \mu_{pei}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{pei} = 0.272 - 0.091 = 0.181$$



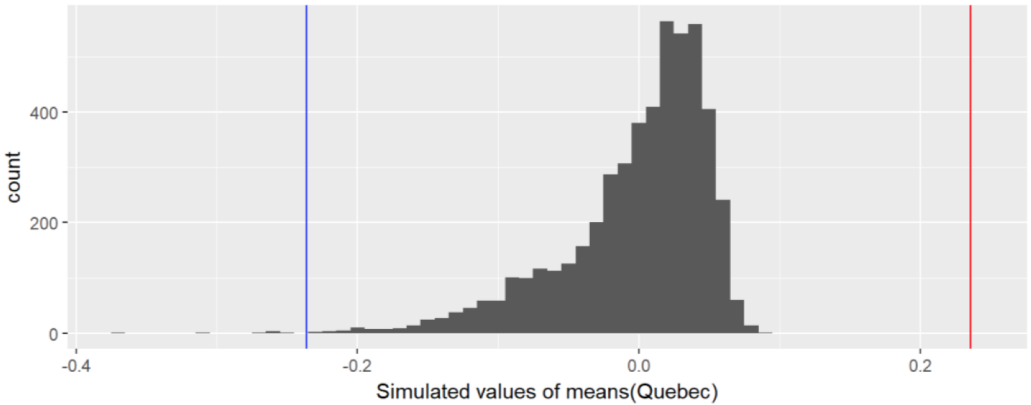
- *P-value = 0.4236*

Sampling Distribution 8:

Newfoundland Vs. Quebec

$$H_0 : \mu_{nfl} = \mu_{qc} \qquad H_a : \mu_{nfl} \neq \mu_{qc}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{qc} = 0.272 - 0.036 = 0.236$$



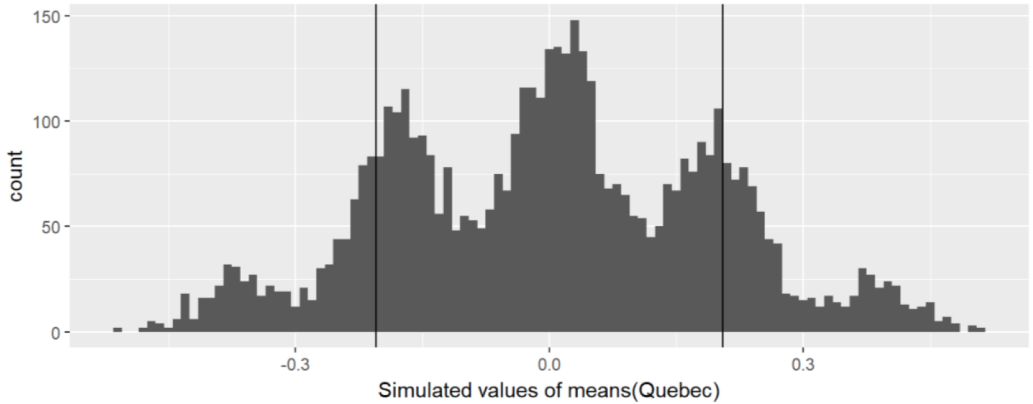
- *P-value = 0.0014*

Sampling Distribution 9:

Newfoundland Vs. Saskatchewan

$$H_0 : \mu_{nfl} = \mu_{sk} \qquad H_a : \mu_{nfl} \neq \mu_{sk}$$

$$\text{Test statistic} = \hat{\mu}_{nfl} - \hat{\mu}_{sk} = 0.272 - 0.067 = 0.205$$



- *P-value = 0.2976*

Results: P-value of Newfoundland v.s Other Provinces

	province	p_value
1	Alberta	0.0144
2	British Columbia	0.0000
3	Manitoba	0.0000
4	New Brunswick	0.0347
5	Nova Scotia	0.0000
6	Ontario	0.0006
7	Prince Edward Island	0.4236
8	Quebec	0.0014
9	Saskatchewan	0.2976

They all have p-value less than 0.05, which provides strong evidence against null hypothesis, EXCEPT FOR:

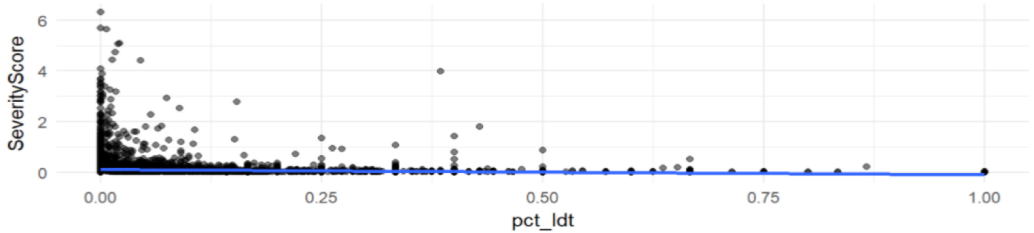
1) *Prince Edward Island* has a comparative small sample size with only 2 observations, so it may explain why it has a p-value of 0.4236.
2) *Saskatchewan* has a p-value of 0.2976, yet it has many high extreme values which does not accurately represent the mean of the whole province.

Regression Model 1:

Relationship Between Percentage of Light Duty Truck and Number of Incidents

- $\beta_0 = 0.111 \quad \beta_1 = -0.206$
- Simple linear regression: $y_1 = 0.111 - 0.206x_1$

```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)  0.1113539 0.003340524  33.33427 2.133848e-231
## pct_ldt     -0.2063132 0.018263146 -11.29670  2.033652e-29
```

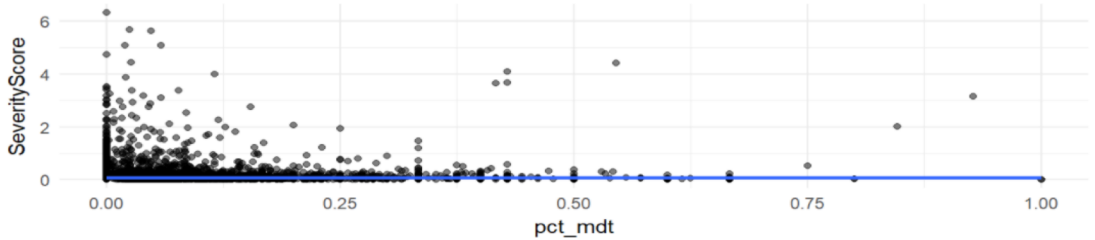


Regression Model 2:

Relationship Between Percentage of Medium Duty Truck and Number of Incidents

- $\beta_0 = 0.0194 \quad \beta_1 = -0.00298$
- Simple linear regression: $y_2 = 0.0194 - 0.00298x_2$

```
##                 Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)  0.091460021 0.003301147 27.7055285 5.562365e-163
## pct_mdt     -0.002981442 0.022539452 -0.1322766  8.947681e-01
```

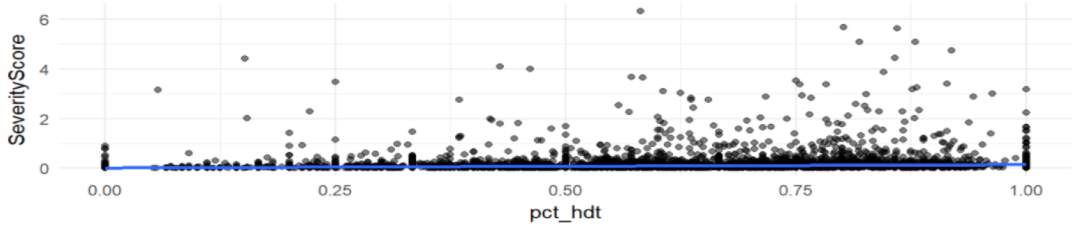


Regression Model 3:

Relationship Between Percentage of Heavy Duty Truck and Number of Incidents

- $\beta_0 = 0.0194 \quad \beta_1 = 0.1326$
- Simple linear regression: $y_3 = 0.0194 + 0.1326x_3$

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 0.01944347 0.006364679   3.054902 2.257129e-03
## pct_hdt     0.13259610 0.010536007 12.585043 4.715212e-36
```



Check if the regression model fit or not:

- By using the coefficient of determination

R^2 from original data set:

[1] 0.1343038

R^2 from using training and test set

[1] 0.109226

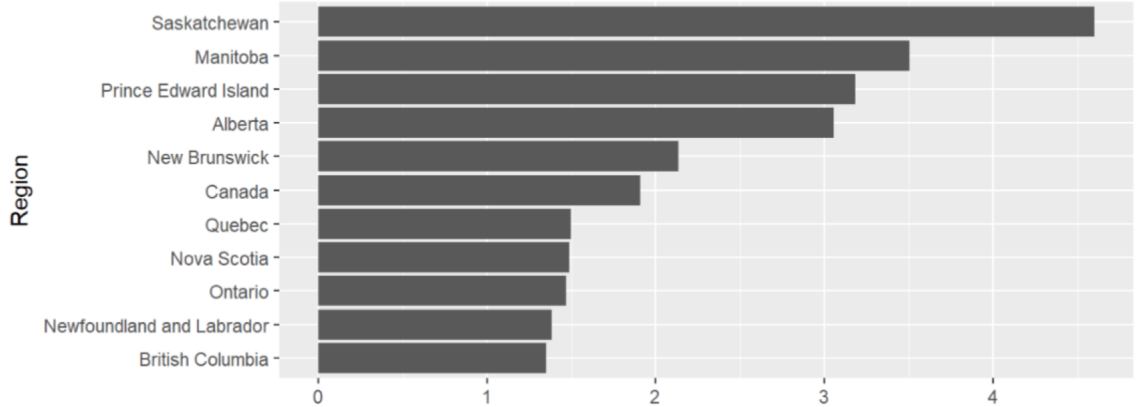
Because R^2 decreases, we decided to add the percentage of each type of trucks as independent factors. Then, our new R^2 becomes:

[1] 0.1448378

- Which lead us to the conclusion that truck type is an insignificant factor to level of hazardous

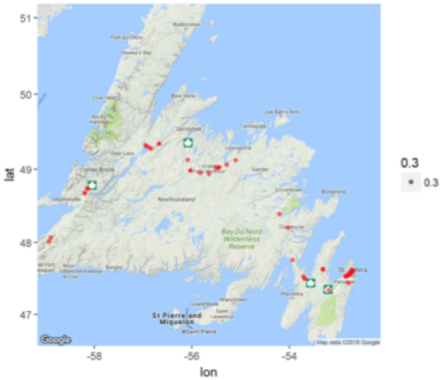
Data From Statistic Canada

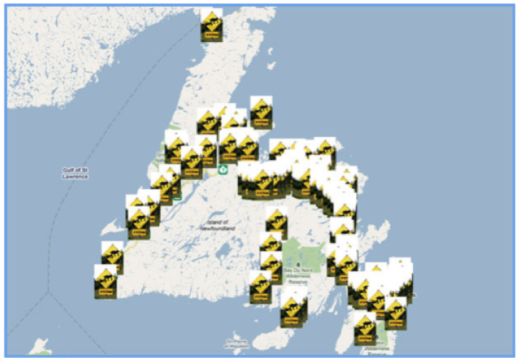
- The imported database contains vehicle registrations statistics in Canada.
- **Variable used**: *Total vehicle registrations* and *registered Heavy duty truck*.



- **Purpose** is to compare the percentage of registered heavy duty trucks in each province.
- **Finding**: Saskatchewan has the most percentage, which is approximatly 4.6%.
- **Possible bias**: the imported database may not be comparable with the databased provided by GeoTab.
- **Methods**: **filter()** interested information, created a **data_frame()** and used **geom_bar()**.

Focus on Newfoundland





If heavy duty trucks do not contribute much to its high SeverityScore, then what may be the possible reason?

- Accidents in Newfoundland seem to mainly occur on highways, where exists many wild mooses.
- Save Our People Action Committee: "*On average there are over 700 moose vehicle accidents in the province*".
- Statistics Canada: "*Newfoundland animal-vehicle collisions are mostly moose-vehicle collisions*".
- Methods: **get_map(), ggmap(), geom_point()** to create map, and **knitr()** to import web image.

Conclusion

- Newfoundland is the most hazardous province with the highest median SeverityScore (Hypothesis Test)
- Low R^2 shows that severity score cannot be predicted by truck type nor by the percentage of Hdtincident
- Imported data of moose accidents may explains the high severity score in NFL

Challenges and Limitations:

- Linear regression model is limited to linear correlations and it's sensitive to outliers.
- More suitable data will be traffic flow of truck instead of truck registration number.
- Could not find the exact number of incidents on highway nor the number of moose vehicle collision.