Careful While Driving in These Provinces!!

Chien-Che Hung(**1004330164**), Yiwen(Leo) Yang (**1004244800**), Jincheng Leng (**1003832695**), Tuoyue Huang (**1003906712**), STA130 TUT0101F,Group F1

Introduction

The dataset from GEOTAB provides the hazardous areas for driving within a specific area. Each area has a severity score that is measured based on harsh braking incidents, traffic flow, and accident-level incidents.

- What's the definition of hazardous driving?
- Which province has the most driving hazards?
- Is the difference between the highest hazardous driving province and other provinces occur by chance?
- Whether the severity score and some subcategories of incidents have relationships?

Statistical Methods

- We aimed to find the province that has the most hazardous driving using **Data Wrangling** and compare it to other provinces using **Barplot**.Barplot can help us to show the difference graphically.
- We found the most dangerous province from the largest proportion of hazardous driving and use **hypothesis tests for two proportions** to show the difference between that province and remaining provinces.
- We attempted to use **linear regression line** to find the relationship between **SeverityScore** and **Heavy Duty Truck**.

Data Summary

- Created **hazardcanada** by filtering (**filter()**) **Canada** from **hazarddat**
- Use **quantile()** to find the **SeverityScore** at position 70% –> 0.0542
- SeverityScore >= 0.0542: **Yes**(This is our definition) SeverityScore < 0.0542: **No**
- Use **mutate()** and **ifelse()** to create a new variable **hazardous_or_not** that contains **Yes** and **No** in a new data frame **hazardcanadawith_def**

	SeverityScore	State	hazardous_or_not	Country
1	0.1155	Alberta	Yes	Canada
2	0.4326	Alberta	Yes	Canada
3	0.1731	Alberta	Yes	Canada
4	0.0428	Alberta	No	Canada
5	0.2321	Alberta	Yes	Canada
6	0.0762	Alberta	Yes	Canada

Province with the most driving hazards

- Create a data frame **new** with three variables indicating the name of province(**State**), the number of hazardous driving in the specific province (**num_of_hzds**), and the total number of observations in that province(**num_of_all**)
- Function used to create dataframe **new**: **group_by()**, **count()**, **data.frame()**, and **indexing**
- Create a data frame **final** from **new** by forming a new variable **pct_of_hzds** which is the quotient of the number of hazardous_driving and total number of observation in the province.
- The highest proportion is found by using **arrange(desc())** from **final** and create a new data frame **final_more**

Hypothesis Test for Two Proportions

$$H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 \neq 0$$

- Repeatedly show the proportion differences between the highest province and the remaining several provinces in addition to the second highest province.
- Simulate the calculation of the difference between the highest province and the chosen province for 1000 times by using **for loop**
- After graphing out the distribution, find the **p-value** by calculating the probability of seeing a difference that is as extreme or more extreme than the test statistics we calculated, assuming that there is no difference between the two provinces.

Linear Regression Line

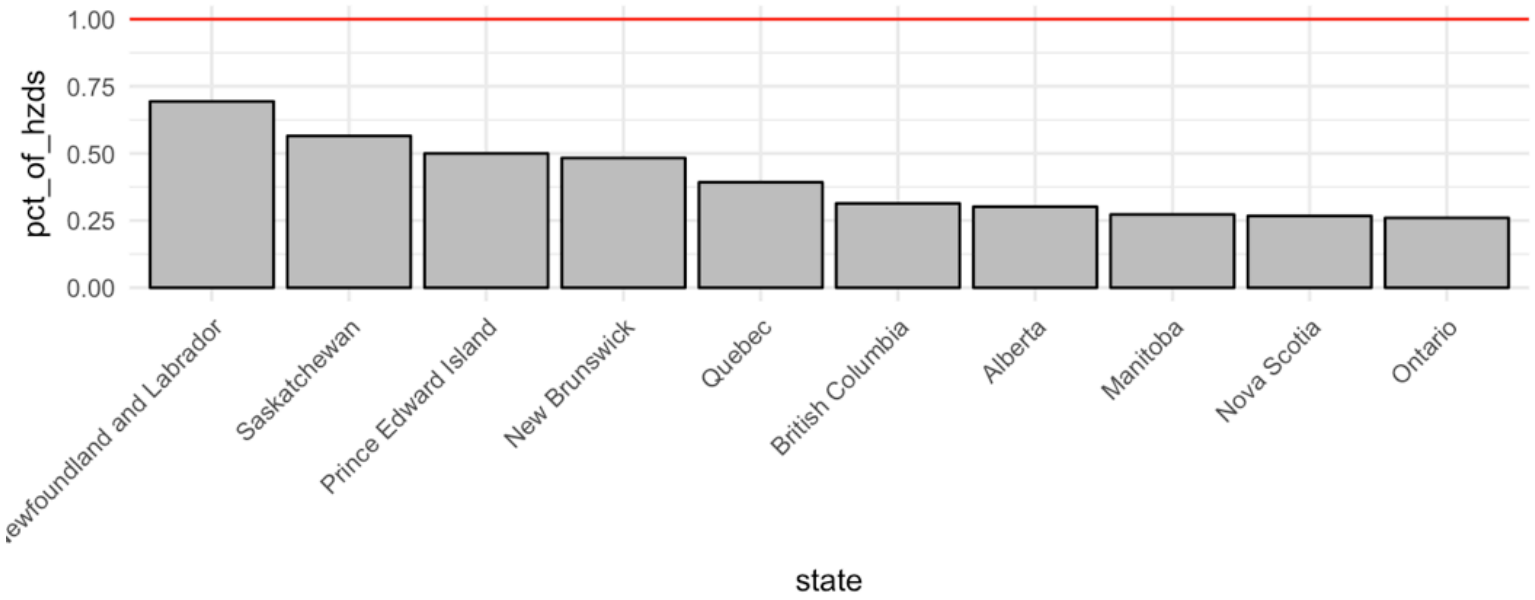
- At the first glance of the data, **hdtIncidents** has large numbers with respect to other incidents
- Create a new dataframe **hazard_variable** from **hazardcanada** by creating **hdt_percentage** (quotient of **HdtIncidents** and **NumberIncidents**)
- Use **ggplot()** by having **x = hdt_percentage** and **y = SeverityScore**
- Use **summary(lm())** to get the **R square** to determine the fitness of the graph and the **p-values** to determine whether two variables have relationships

Results

According to the data, we can see **Newfoundland and Labrador** has the highest percentage of hazardous driving, which indicates it has the most driving hazards.

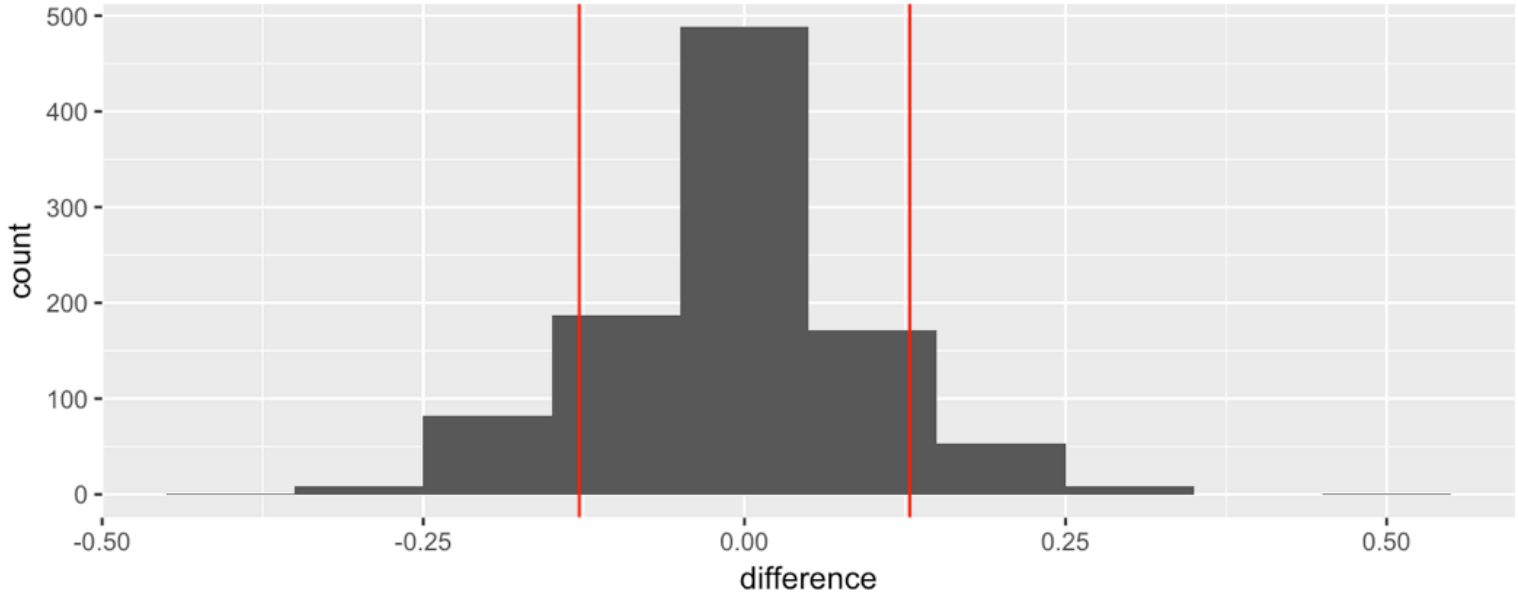
	state	num_of_hzds	num_of_all	pct_of_hzds
1	Newfoundland and Labrador	34	49	0.6938776
2	Saskatchewan	26	46	0.5652174
3	Prince Edward Island	1	2	0.5000000
4	New Brunswick	58	120	0.4833333
5	Quebec	836	2128	0.3928571
6	British Columbia	209	666	0.3138138
7	Alberta	144	477	0.3018868
8	Manitoba	164	601	0.2728785
9	Nova Scotia	65	243	0.2674897
10	Ontario	1538	5909	0.2602809

Bar Plot from dataset **final_more**



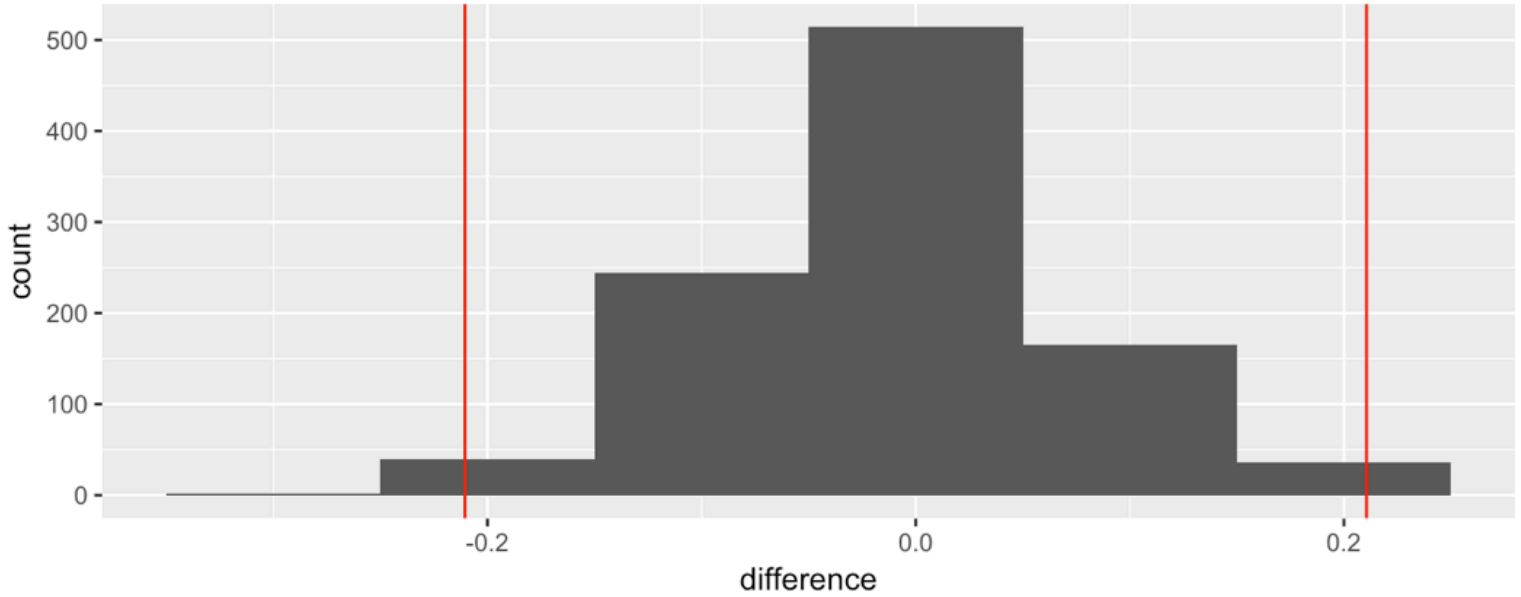
Hypothesis Test for **Newfoundland and Labrador** vs. **Saskatchewan**

```
  p_value
1   0.234
```



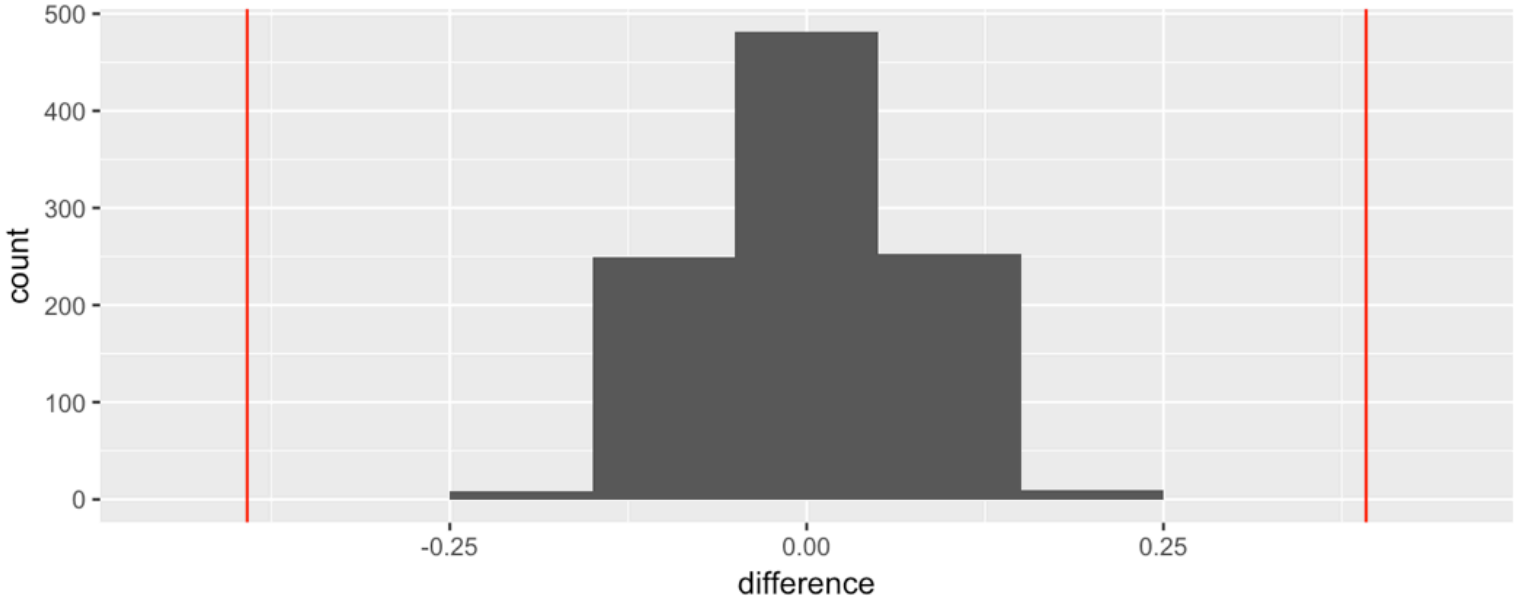
Hypothesis Test for **NewFoundland and Labrador** vs. **New Brunswick**

```
  p_value
1   0.013
```



Hypothesis Test for **Newfoundland and Labrador** vs. **Alberta**

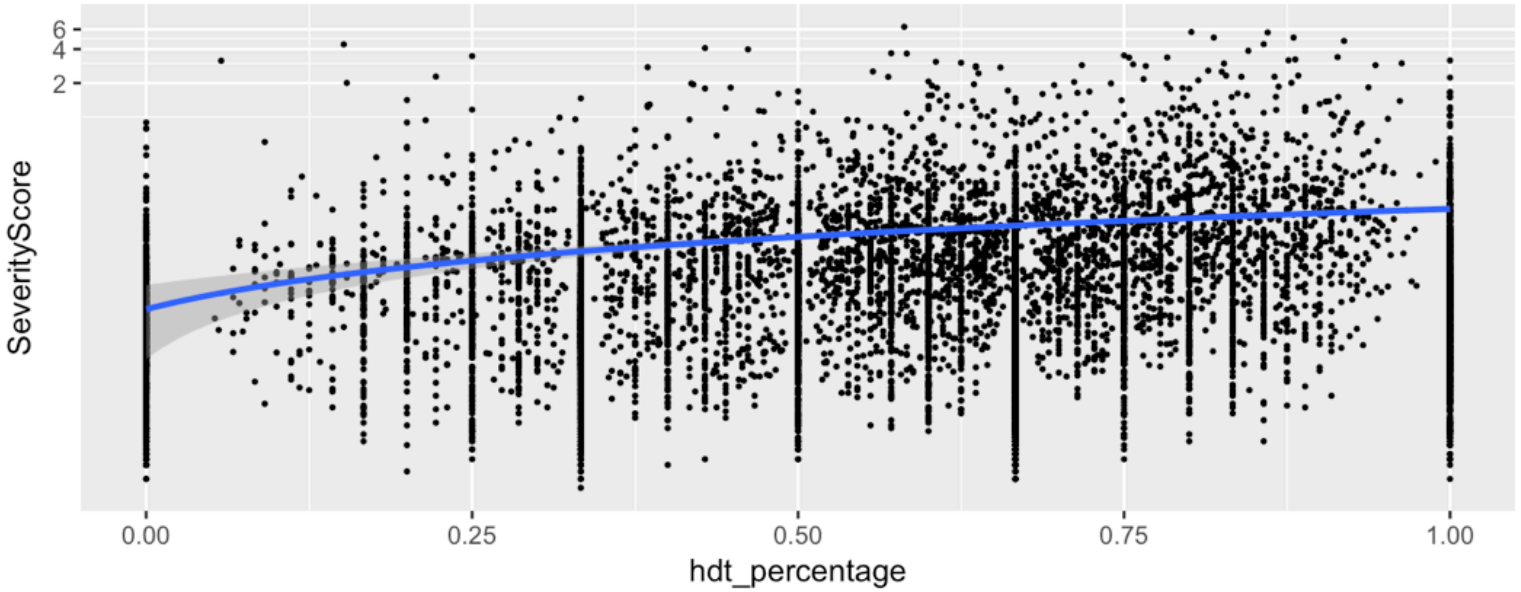
```
  p_value
1       0
```



Linear regression line for **Severity Score** and **hdt_percentage** with **log transfermation**

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

```
[1] 0.015233
```



Conclusion

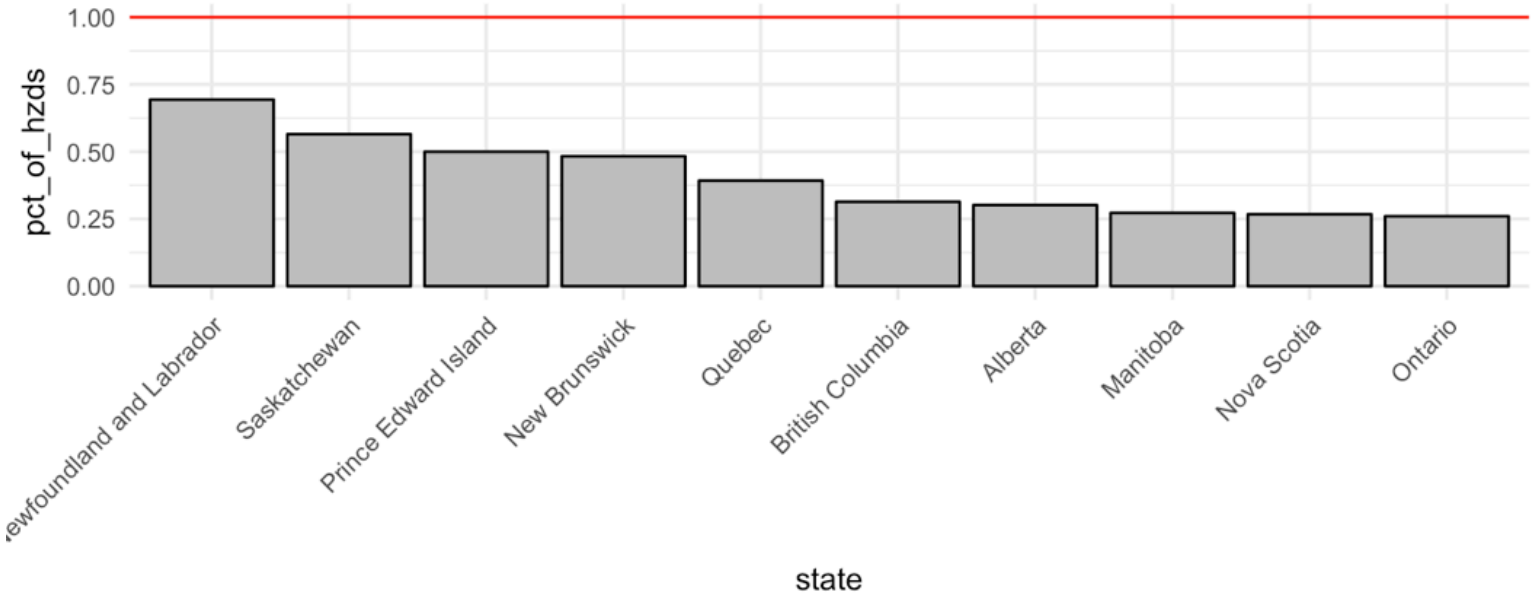
Definition of Hazardous Driving: When the severity score exceeds 0.0542, we consider it as hazardous driving.

Three questions:

- Which province has the most driving hazards?
- Is the difference between the highest hazardous driving province and other provinces occur by chance?
- Whether the severity score and some subcategories of incidents have relationships?

Which province has the most driving hazards?

From the bar plot we created , we are able to see the sequence clearly since it is arranged in a descending order. Newfoundland and Labrador is definitely the most dangerous place for driving according to our definition. Saskatchewan follows Newfoundland and Labrador while Ontario is the safest place to drive in.



What are the relationships betweeen the province with most driving hazards and other provinces?

- We do not have sufficient evidence to say that there is a difference in hazardous driving between **Newfoundland and Labrador** and **Saskatchewan**.
- We have moderate evidence to say that there exists a difference in hazardous driving between **Newfoundland and Labrador** and **New Brunswick**.
- We have very strong evidence to say that there exists a difference in hazardous driving between **Newfoundland and Labrador** and **Alberta**.

Provinces	Saskatchewan	New Brunswick	Alberta
Newfoundland and Labrador	p_value = 0.234	p_value = 0.013	p_value = 0

Whether the severity score and some subcategories of incidents have relationships?

- It is not a good estimate of relationship between **hdt_percentage** and **severity score** since the R square = 0.1523 which represents a poor fit.
- Assuming the slope of regression line is zero, the **p_value is 4.715212e-36**, indicating that we have very **strong evidence** to against that the **slope is 0**.
- It is hard to tell the relationship between the two variables since the line is actually a poor fit.

Through performing numerous data wrangling, cleaning on the data given, we conclude that **Newfoundland and Labrador** is the most dangerous province to drive in according to our definition.

Challenge:

- How can we accurately define hazardous drivng?
- What is the most suitable approach?
- What can we do for the neglected data?