STAA57 Group Project

What are the dominant research areas in Ontario's automotive sector, and how do institutions specialize in different fields

Minh Tran - 1006804914, Swajeet Jadhav - 1009888276

2025-03-15

I. Introduction

Research Goal

With the growing popularity of electric vehicles (EVs) and the rapid advancements in artificial intelligence (AI), particularly in the development of autonomous vehicles like Tesla, it is crucial for automotive research facilities in Ontario to focus their efforts and resources on the right areas. This will ensure a significant contribution to the academia world while assisting the government in implementing effective policies and educational institutions better educate, prepare the next generation of labors.

This report aims to examine the automotive research areas prioritizing by Ontario's research facilities and identify specialized facilities in trending fields. It will also explore which areas are attracting the most researchers and which institutions are specializing in particular research domains.

Dataset description

Data source

Our data set was collected by researchers from the Ministry of Economic Development, Job creation and Trade and was published on the Government of Ontario's public database in 2018. A supporting description table of the research tag will be provided in the Appendix for better understanding of the main data set values.

Data set Variables

```
## Rows: 545
## Columns: 37
## $ Institution                    <chr> "Brock University", "Carleton Univers~
## $ Researcher.Name                <chr> "Ahmed, Syed Ejaz", "Ahmadi, Mojtaba"~
## $ Associated.Facilities          <chr> "Centre for Statistical Consulting", ~
## $ Research.Areas                 <chr> "Traffic and road injury prevention."~
## $ Research.Chairs.Grant.Funding  <chr> "", "", "", "", "Canada Research Chai~
## $ Tag.1                          <chr> "Injury Prevention", "Mechatronics", ~
## $ Tag.2                          <chr> "", "Control", "Autonomy and AI", "Ne~
## $ Tag.3                          <chr> "", "", "Sensors", "", "", "", "", ""~
## $ Tag.4                          <chr> "", "", "", "", "", "", "", "", "", "~
## $ Tag.5                          <chr> "", "", "", "", "", "", "", "", "", "~
## $ Alternative.Fuels              <chr> "", "", "", "", "", "", "", "", "", "~
## $ Autonomy.and.AI                <chr> "", "", "x", "", "", "", "x", "", "",~
## $ Batteries.and.Fuel.Cells       <chr> "", "", "", "", "", "", "", "", "", "~
## $ Biocomposites                  <chr> "", "", "", "", "", "", "", "", "", "~
## $ Coatings.and.Corrosion         <chr> "", "", "", "", "", "", "", "", "", "~
## $ Connected.Vehicles             <chr> "", "", "", "", "", "", "", "", "", "~
## $ Control                        <chr> "", "x", "x", "", "", "", "", "", "",~
## $ Crashworthiness                <chr> "", "", "", "", "", "", "", "", "", "~
## $ Electronics                    <chr> "", "", "", "", "", "", "", "", "", "~
## $ Forming.and.Joining            <chr> "", "", "", "", "", "", "", "", "", "~
## $ High.Strength.Steel            <chr> "", "", "", "", "", "", "", "", "", "~
## $ Hybrid.and.Electric.Vehicles   <chr> "", "", "", "", "", "", "", "", "", "~
## $ Industrial.Processes           <chr> "", "", "", "", "", "", "", "", "", "~
## $ Injury.Prevention              <chr> "x", "", "", "", "", "", "", "", "", ~
## $ Internal.Combustion.Engines    <chr> "", "", "", "", "", "", "", "", "", "~
## $ Lightweight.Metals             <chr> "", "", "", "", "", "", "", "", "", "~
## $ Mechatronics                   <chr> "", "x", "", "", "", "", "", "", "", ~
## $ Nanotechnology                 <chr> "", "", "", "", "", "", "", "", "", "~
## $ Networks.and.Security          <chr> "", "", "", "x", "", "x", "", "x", "x~
```

```
## $ Noise..Vibration.and.Harshness   <chr> "Err:507", "Err:507", "Err:507", "Err~
## $ Polymers.and.Composite.Materials <chr> "", "", "", "", "", "", "", "", "", "~
## $ Powertrain                       <chr> "", "", "", "", "", "", "", "", "", "~
## $ Sensors                          <chr> "", "", "x", "", "x", "", "x", "", ""~
## $ Software                         <chr> "", "", "", "x", "", "", "", "", "", ~
## $ Stress.and.Fracture              <chr> "", "", "", "", "", "", "", "", "", "~
## $ Transportation.and.Charging      <chr> "", "", "", "", "", "", "", "", "", "~
## $ Vehicle.Design                   <chr> "", "", "", "", "", "", "", "", "", "~
```

Based on a summary of the data set, the `Noise..Viration.and.Harshness` variable was corrupted during the process of uploading or collecting the data and therefore will not be used in our report

Some of the main categorical variables that can help categorizing the data include

- **Researcher.Name**: Names of the researchers.
- **Institution**: Name of university or research center that the researcher works at
- **Research.Areas**: Researcher general field of research
- **Tag**: Researcher's specialized field of research

In addition, the data set also contains research fields as variable (e.g. Alternative Fuels, Autonomy and AI, Vehicle Design) to indicate whether or not a researcher work fall into that field of research

2. Data Overview Analysis (Swajeet)

2.1. Descriptive Statistic

Tables

An initial Summary table of all research areas.

Table 1: Summary statistics of research areas

Research_Area	Count	Frequency
Networks and security	92	0.0795848
Autonomy and AI	79	0.0683391
Transportation and charging	65	0.0562284
Polymers and composite materials	64	0.0553633
Lightweight metals	61	0.0527682
Batteries and fuel cells	59	0.0510381
Forming and joining	57	0.0493080
Industrial processes	57	0.0493080
Connected vehicles	55	0.0475779
Nanotechnology	52	0.0449827
Sensors	48	0.0415225
Hybrid and electric vehicles	47	0.0406574
Software	47	0.0406574
Vehicle design	37	0.0320069
Control	35	0.0302768
Injury Prevention	35	0.0302768
Electronics	32	0.0276817
Alternative fuels	31	0.0268166
Biocomposites	30	0.0259516
Coatings and corrosion	30	0.0259516
Internal combustion engines	28	0.0242215

Research_Area	Count	Frequency
Powertrain	27	0.0233564
Stress and fracture	27	0.0233564
Mechatronics	17	0.0147059
High strength steel	15	0.0129758
Noise, vibration and harshness	13	0.0112457
Crashworthiness	12	0.0103806
Other	4	0.0034602

The table above presents a summary statistic of research areas, indicating the number of researchers engaged in each category (Count) and their relative proportions ("Frequency"). The results reveal that **Networks and Security**, **Autonomy and AI**, and **Transportation and Charging** are the most prominent fields, attracting the highest number of researchers

To further example institutional alignment with these research trends, the data set will be analyzed to identify the top five institutions employing the largest number of researchers in each of the three leading fields identified above.

Table 2: Top 5 Institutions in Networks and security

Institution	Reseacher_count
Carleton University	37
University of Waterloo	22
University of Ontario Institute of Technology (UOIT)	7
Ryerson University	6
University of Toronto	6

Table 3: Top 5 Institutions in Autonomy and AI

Institution	Reseacher_count
Carleton University	23
University of Waterloo	16
University of Toronto	11
Ryerson University	6
University of Windsor	5

Table 4: Top 5 Institutions in Transportation and charging

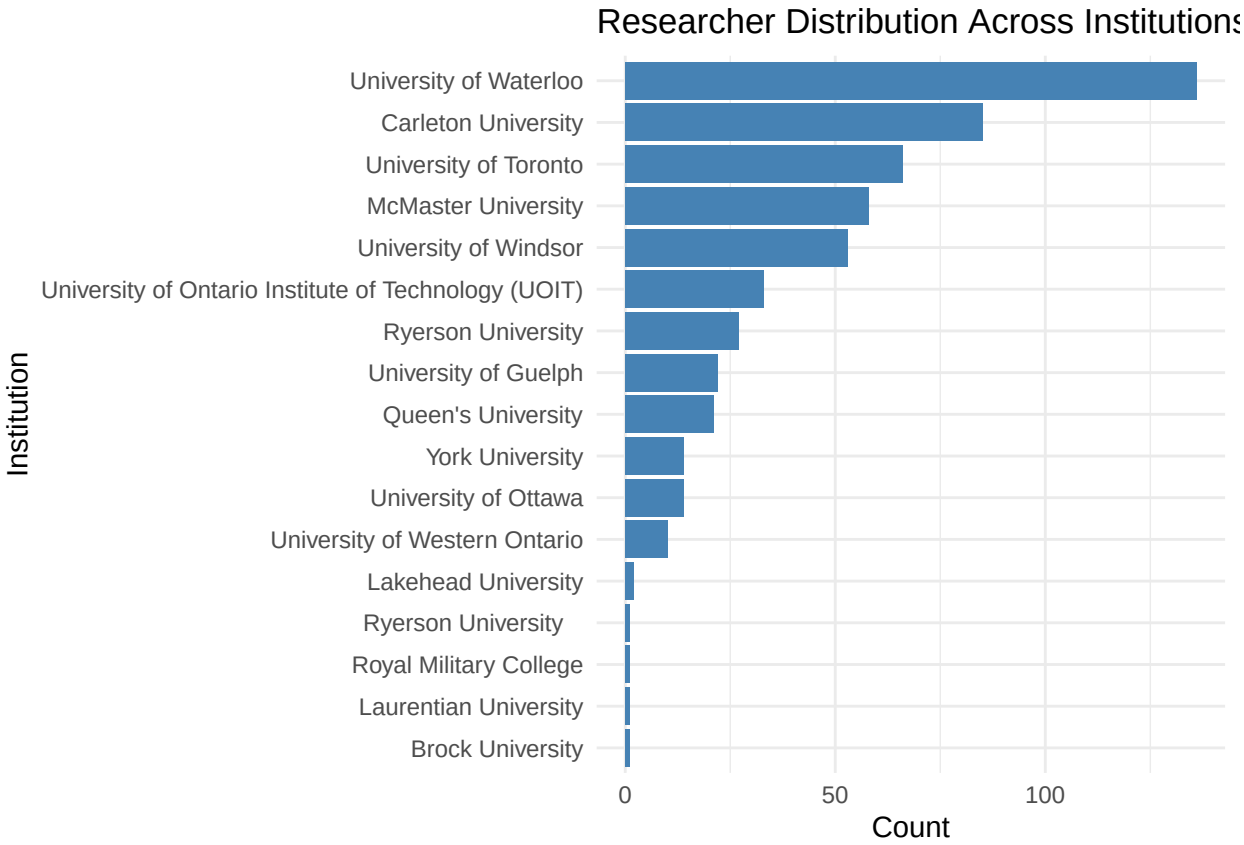
Institution	Reseacher_count
Carleton University	13
University of Waterloo	12
Ryerson University	11
University of Toronto	6
York University	6

Carleton University has a significance number of researchers across three leading research fields, followed closely by the University of Toronto, University of Waterloo and Ryerson University (now Toronto Metropolitan University). This trend aligns with the fact that these institutions are large, metropolitan universities,

particularly in cities such as Toronto and Ottawa, which have established reputations for excellence in technological research. Given their substantial resources and expertise, these universities play a crucial role in shaping Canada's research priorities and should be central to discussions on nation's research agenda

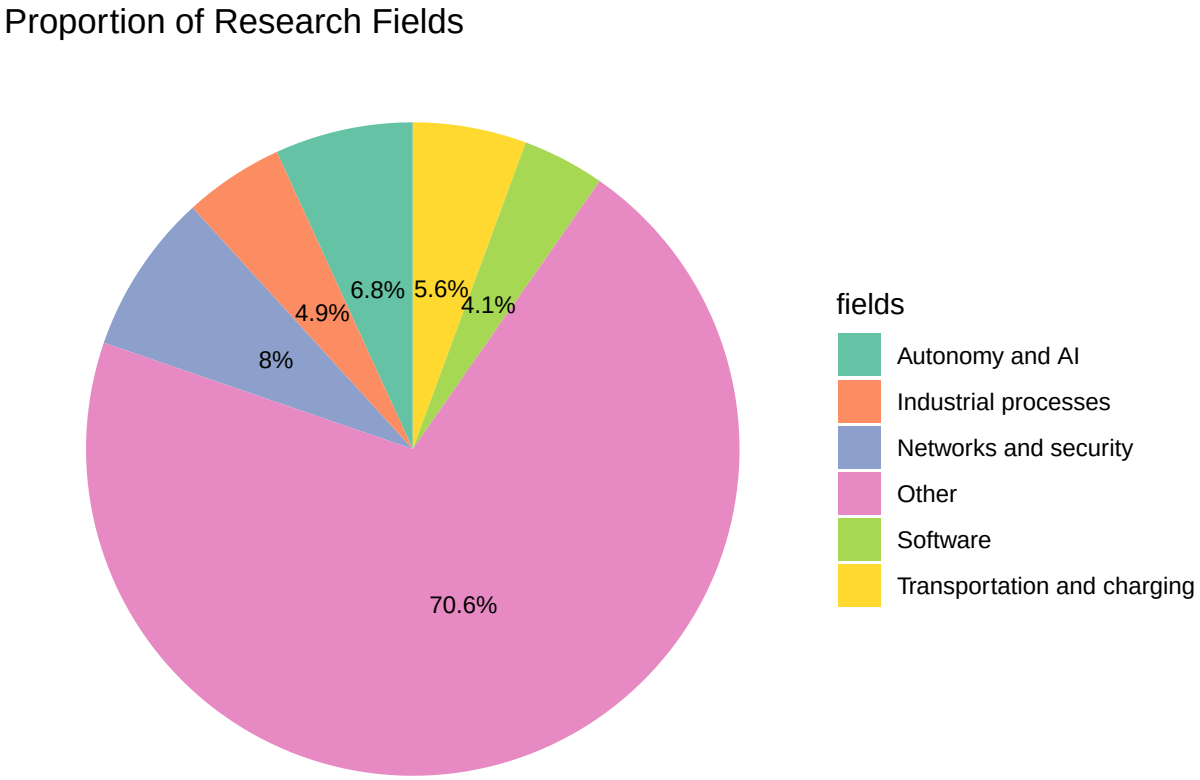
2.2. Graph & Visualizations

2.2.1 An informative Bar chart on Researchers Distribution



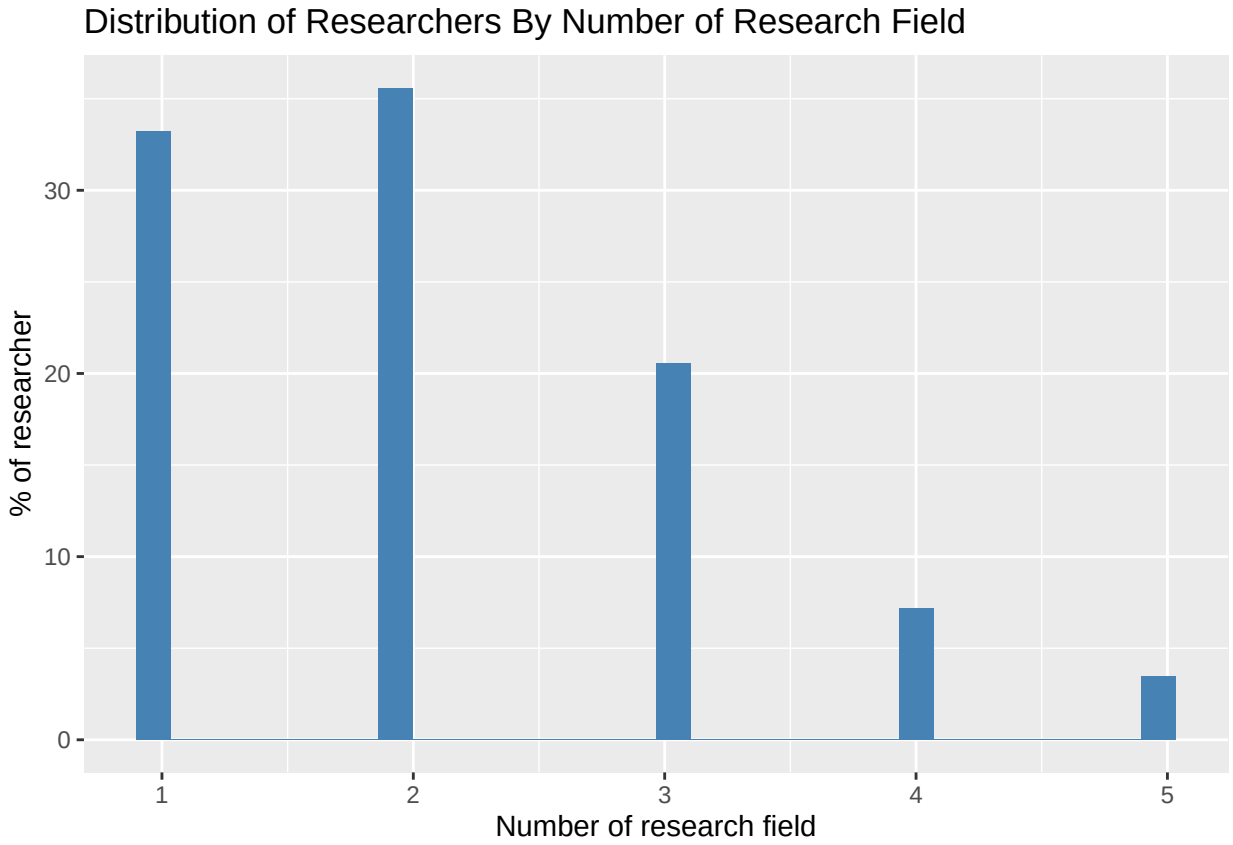
The graph presents a visual representation of total number of researchers employed by each institutions. The findings reinforce our conclusions drawn in Section 2.1, confirming that leading institutions such as University of Waterloo, Carleton University, University of Toronto continue to have the highest number of researchers engaged with the automotive industry. Notably while Carleton University leads in term of researchers specializing in key fields, University of Waterloo has the largest number of researchers overall. This data may serve as a valuable resource for policy makers and investors, enabaling them to strategically allocate support to institutions that have the potential to contribute significantly to automotive research.

2.2.2 An interesting Pie Chart on Research Fields



Each segment of the graph represents the portion of researchers engaged in a specific field, based on data aggregated from multiple institutions. Given the number of research categories, the visualization focuses on the five fields with the highest concentration of researchers, while grouping the remaining fields under the category "Other". The data reveal that although the top five fields account for the majority of researcher, the differences among them is not significant. Furthermore as we can see that the Other group still maintain the significant portion of researchers indicating that fields in the top 5 group are not distinct from fields in the Other group by a significant gap.

2.2.3 Histogram of Research Variance by Tags.



This diagram illustrates the distribution of researchers based on the number of research fields they are associated with, as indicated by the "Tag" columns. If a researcher has a value in only one "Tag" column, they are considered to be specialized in a single field. Similarly, researchers with values in multiple columns are associated with multiple fields. The chart results indicate that the majority of researchers in the data set specialize in only one or two research fields as these two categories collectively account for more than 60% of the total researcher population.These findings could serve as a valuable resource for universities and educators, enabling them to tailor their efforts toward training researchers with specialized or generalized expertise according to the needs of the industry.

III. Statistical Analysis (rayaanxsyed)

We could pick an area we think that will be most trending (AI for example) and test to see if that is true

- Confidence interval for the average number of researcher of all data set institutions

```
## [1] 12.98446 51.13318
## attr(,"conf.level")
## [1] 0.95
```

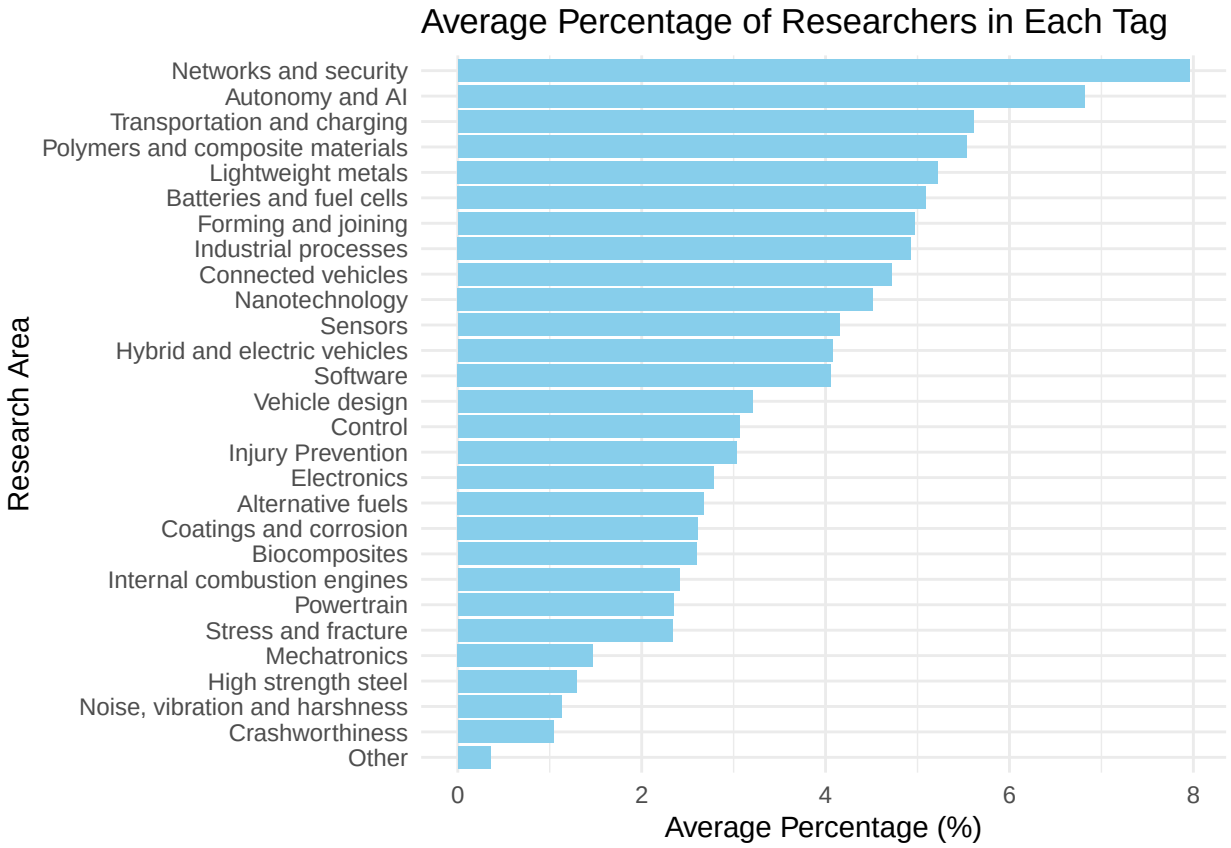
- Let's choose the most trending research topic (Network and security) and do a confidence interval for the average number of researchers in the field.

```
## [1]  0.3903509 10.4331786
## attr(,"conf.level")
## [1] 0.95
```

Hypothesis Testing For the topic (Network and security) and do a hypothesis test for the topic.

```
##
##  One Sample t-test
##
## data:  network_security_count
## t = 0.17384, df = 16, p-value = 0.8642
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##   0.3903509 10.4331786
## sample estimates:
## mean of x
##  5.411765
```

Let's pick our highest leading tag and see how it performs in bootstrapped data. We want to observe how it does compared to other tags and if it's the leading tag in all the samples by takign an average.



In this graph, the data of the institutions and their tags have been replicated 1000 times. We take the average percentage to see how they perform over the 1000 trials and gain a graph of each tags performance.

```
## 95% Confidence Interval for Network and Security: [ 6.401384 ,  9.515571 ]
```

By taking a confidence interval for Network and Security, we observe that the true percentage of researchers in Network and Security lies in the interval [6.468531, 9.702797]. This is a very high value when comparing this to the other tags in our horizontal bar chart.

Let's examine how Networks and Security performs when compared to the 2nd highest tag (Autonomy and AI) in our bootstrapped data to see if it always leads the data by a hypothesis test.

```
## Paired t-test: Networks and Security vs. Autonomy and AI
```

```
## Mean % Network and Security: 7.958304
```

```
## Mean % ( Autonomy and AI ): 6.818426
```

```
## T-statistic: 32.61985
```

```
## P-value: 8.496615e-160
```

```
## Reject H . Networks and Security is significantly higher than the other tags. (p < 0.05).
```

Based off our paired t-test, we get a average Network Security score of 8.001923 and a average Autonomy and AI score of 6.806818. Since the averages are distant, we can expect a large t statistic and a small p value.

We observe a t statistic value of 31.29503 and a extremely small p value 1.070572e-150 ($p < 0.05$), rejecting the null hypothesis and observing Networks and Security is significantly higher than the other tags.

- Bootstrapping is a re sampling technique used to estimate statistics on a dataset by repeatedly sampling with replacement. It helps assess the variability of an estimator (e.g., mean, standard deviation) and

construct confidence intervals without relying on strong parametric assumptions. Do bootstrapping on network and security.

```
##      2.5%     97.5%
##  1.882353 10.355882
```

```
##
##  One Sample t-test
##
## data:  network_security_count
## t = 2.2847, df = 16, p-value = 0.03632
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   0.3903509 10.4331786
## sample estimates:
## mean of x
##  5.411765
```

IV. Predictive modeling (Regression) (Minh)

1. Logistic Regression Model

1.1. Model explanation

The allocation of research funding plays a pivotal role in identifying key research fields that attract the attention of funding bodies such as Canada Research Chair Program and other major investors. These funding trend are essential indicators of the areas prioritized by the Canadian government and industry stakeholders in the automotive sector. Understanding these trends can provide insight into the direction of research investment, thereby influencing strategic decision in academia and industry alike.

To analyze the factors influencing research funding, we employed a logistic regression model with the primary dependent variable being **is_Funded**. This variable take a value of 1 if the research has secured funding from the Canada Research Chair program or other similar grants, and 0 otherwise. The independent variables include various research's primary fields as categorized in **Tag.1** column of the data set.

1.2. Result and Key Findings

Table : Regression Model Result Summary

```
##
## Call:
## glm(formula = is_Funded ~ ., family = binomial, data = reg_data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.18232    0.60553  -0.301   0.7633
## is_Alternative.Fuels              1.43508    0.82951   1.730   0.0836 .
## is_Autonomy.and.AI               -0.04082    0.67577  -0.060   0.9518
## is_Batteries.and.Fuel.Cells       0.65233    0.72744   0.897   0.3699
## is_Biocomposites                  0.40547    0.76920   0.527   0.5981
## is_Coatings.and.Corrosion         0.02817    0.82231   0.034   0.9727
## is_Connected.Vehicles             0.81093    0.74722   1.085   0.2778
## is_Control                       -1.11696    0.88933  -1.256   0.2091
## is_Crashworthiness                1.28093    1.30384   0.982   0.3259
## is_Electronics                    0.43364    0.78780   0.550   0.5820
## is_Forming.and.Joining            1.62924    0.82188   1.982   0.0474 *
## is_High.Strength.Steel           -0.22314    1.09545  -0.204   0.8386
```

```
## is_Hybrid.and.Electric.Vehicles     -0.04082    0.90370  -0.045   0.9640
## is_Industrial.Processes              0.30010    0.69622   0.431   0.6664
## is_Injury.Prevention                 0.51879    0.84233   0.616   0.5380
## is_Internal.Combustion.Engines       1.38629    0.89443   1.550   0.1212
## is_Lightweight.Metals                0.77011    0.72265   1.066   0.2866
## is_Mechatronics                     -0.10536    0.97468  -0.108   0.9139
## is_Nanotechnology                    0.69315    0.79582   0.871   0.3838
## is_Networks.and.Security            -0.53330    0.65905  -0.809   0.4184
## is_Polymers.and.Composite.Materials  0.61764    0.71861   0.859   0.3901
## is_Powertrain                        0.18232    1.01653   0.179   0.8577
## is_Sensors                          -0.40547    0.82327  -0.493   0.6224
## is_Software                         -0.46431    0.71077  -0.653   0.5136
## is_Stress.and.Fracture               0.87547    1.05672   0.828   0.4074
## is_Transportation.and.Charging      -0.35667    0.69265  -0.515   0.6066
## is_Vehicle.Design                    0.47000    0.74907   0.627   0.5304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 755.31  on 544  degrees of freedom
## Residual deviance: 704.64  on 518  degrees of freedom
## AIC: 758.64
##
## Number of Fisher Scoring iterations: 4
```

Table : Regression Model Coefficient

	coefficient	odd
is_Forming.and.Joining	1.6292405	5.10
is_Alternative.Fuels	1.4350845	4.20
is_Internal.Combustion.Engines	1.3862944	4.00
is_Crashworthiness	1.2809338	3.60
is_Stress.and.Fracture	0.8754687	2.40
is_Connected.Vehicles	0.8109302	2.25
is_Lightweight.Metals	0.7701082	2.16
is_Nanotechnology	0.6931472	2.00
is_Batteries.and.Fuel.Cells	0.6523252	1.92
is_Polymers.and.Composite.Materials	0.6176396	1.86
is_Injury.Prevention	0.5187938	1.69
is_Vehicle.Design	0.4700036	1.61
is_Electronics	0.4336360	1.55
is_Biocomposites	0.4054651	1.50
is_Industrial.Processes	0.3001046	1.35
is_Powertrain	0.1823216	1.20
is_Coatings.and.Corrosion	0.0281709	1.03
is_Autonomy.and.AI	-0.0408220	0.96
is_Hybrid.and.Electric.Vehicles	-0.0408220	0.96
is_Mechatronics	-0.1053605	0.90
(Intercept)	-0.1823216	0.84
is_High.Strength.Steel	-0.2231436	0.80
is_Transportation.and.Charging	-0.3566749	0.71
is_Sensors	-0.4054651	0.67
is_Software	-0.4643056	0.63

	coefficient	odd
is_Networks.and.Security	-0.5332985	0.59
is_Control	-1.1169614	0.33

A substantial proportions of the variables int he model exhibit high **p-value** of greater than 5% suggesting that most of the research fields are not statistically significant predictors of whether a research project will be funded. This is likely due to the limited number of observations available int he data set, which may have constrained the model's ability to detect more nuanced relationships. Nevertheless, the model provides valuable insights into general trends in funding allocation by the Canadian government and major investors.

Despite the high **p-value** for many variables, certain research fields stood out interns of their impact on the funding probabilities. Specifically, **Forming and Joining**, **Alternative Fuels**, and **Internal Combustion Engines** demonstrated relatively high log-odds ratios. These fields exhibited 4 to 5 times higher odds of receiving funding than those in other areas. In addition, **is_Forming.and.Joining** and **is_Alternative.Fuels** variables were both statistically significant with p-values of approximately 4% and 10% respectively which further support the findings above.

The significance and high odd of **Forming and Joining** research fields aligns with ongoing research efforts to improve manufacturing process which are crucial for the production of commercial vehicles. While the corresponding number for **Alternative Fuels** underscores the importance of environmentally sustainable technologies, which are increasingly emphasized in both governmental policy and industry innovation.

2) Cross validation

To further evaluate the robustness of the logistic regression model, we conducted a k-fold cross-validation with four folds.

The results are as follow:

```
## AUC score: 0.6168478 0.6171181 0.5707246 0.5953256
```

```
## Average AUC score: 0.600004
```

The model predictive performance assessed using the AUC score. The resulting average AUC score was 0.609, which suggests a fair to somewhat weak performance. This moderate performance is likely attributed to the large number of non-significant variables in the model, reflecting the challenges posed by limited data. Nevertheless, the model remains useful for identifying general funding trends rather than providing highly accurate predictions for individual research projects. It serves its purpose of illustrating that dominant research areas that are attracting funding, which is the primary objective of the analysis.

V. Summary

In conclusion…

The logistic regression analysis provides valuable insights into the research funding landscape within the Canadian automotive sector. The significant fields of **Forming and Joining** and **Alternative Fuels**, along with the notable importance of **Internal Combustion Engines**, underscore the ongoing focus on traditional automotive manufacturing processes and sustainable technologies. These finding aligns with the broader trends observed in the automotive industry, where there is increasing attention to environmentally friend alternatives to conventional fuels and more efficient manufacturing practices.

However, as the market of Autonomous Vehicles and Electric Vehicles grows, we believe that more resources should be reallocate to support researches in these emerging areas. Fields such as **Autonomy and AI**, **Batteries and Fuel Cells**, and **Hybrid and Electric Vehicles** are expected to play a pivotal role in shaping the future of automotive industry. # Appendix