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

1. Introduction

1.1. 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 categories prioritized by Ontario's research facilities and identify specialized facilities in trending fields. It will also explore which areas are attracting the most funding from the Canadian government and major funds. Specifically this study seeks to address teh following questions:

1. What is the leading research category in the automotive industry based on the number of researchers?
2. Which institutions are aligning with research trends by employing a significant number of researchers in the top three leading research categories?
3. Which primary research categories are attracting the most funding

1.2. 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. Each entry record a researcher's information regarding their working place and the research areas of expertise. A supporting table providing better descriptions of the research tags will be provided in the Appendix.

Here are the variables in the data set:

```
##  [1] "Institution"                       "Researcher.Name"
##  [3] "Associated.Facilities"             "Research.Areas"
##  [5] "Research.Chairs.Grant.Funding"     "Tag.1"
##  [7] "Tag.2"                             "Tag.3"
##  [9] "Tag.4"                             "Tag.5"
## [11] "Alternative.Fuels"                 "Autonomy.and.AI"
## [13] "Batteries.and.Fuel.Cells"          "Biocomposites"
## [15] "Coatings.and.Corrosion"            "Connected.Vehicles"
## [17] "Control"                           "Crashworthiness"
## [19] "Electronics"                       "Forming.and.Joining"
## [21] "High.Strength.Steel"               "Hybrid.and.Electric.Vehicles"
## [23] "Industrial.Processes"              "Injury.Prevention"
## [25] "Internal.Combustion.Engines"       "Lightweight.Metals"
## [27] "Mechatronics"                      "Nanotechnology"
## [29] "Networks.and.Security"             "Noise..Vibration.and.Harshness"
## [31] "Polymers.and.Composite.Materials" "Powertrain"
## [33] "Sensors"                           "Software"
## [35] "Stress.and.Fracture"               "Transportation.and.Charging"
## [37] "Vehicle.Design"
```

Some of the most important variable that will mainly be used in this reports are

1. Institution: Name of the researcher institution
2. Researcher.Name: Researcher name
3. Research.Areas: Researcher general research area
4. Researcher.Chairs.Grant.Funding: Name of the funding if that researcher have one
5. Tag. 1 ~ 5: Categorization of researchers research areas
6. Remaining columns: Research categories. If a researchers is part of a category they will have an x in that column.

Data clean up

Based on the data set summary, our team identified that the 'Noise..Vibration.and.Harshness' variable was corrupted during data collection or uploading. As a result, it will not be included in our report. Additionally, we observed that Brock University, Lakehead University, Royal Military College, and Laurentian University each had fewer than two researchers, suggesting incomplete data collection for these institutions. Therefore, researcher entries from these institutions will also be excluded.

II. Data Overview Analysis (Swaaa)

Descriptive Statistic

- Table 2: Summary statistics of research areas (count, frequency)
    - Identify the most common research fields

```r
research_count = researchers %>% group_by(Tag.1) %>%
  summarise(Count = n()) %>%
  mutate(Frequency = Count / sum(Count)) %>%
  arrange(desc(Count))
top_research_field = research_count$Tag.1[1]
kable(research_count, caption = "Summary statistics of research areas")
```

Table 1: Summary statistics of research areas

Tag.1	Count	Frequency
Networks and security	67	0.1240741
Autonomy and AI	44	0.0814815
Transportation and charging	38	0.0703704
Industrial processes	34	0.0629630
Software	32	0.0592593
Lightweight metals	28	0.0518519
Batteries and fuel cells	26	0.0481481
Polymers and composite materials	26	0.0481481
Connected vehicles	23	0.0425926
Forming and joining	21	0.0388889
Vehicle design	21	0.0388889
Alternative fuels	18	0.0333333
Biocomposites	18	0.0333333
Electronics	16	0.0296296
Nanotechnology	16	0.0296296
Control	14	0.0259259

Tag.1	Count	Frequency
Sensors	14	0.0259259
Coatings and corrosion	13	0.0240741
Internal combustion engines	13	0.0240741
Injury Prevention	10	0.0185185
Hybrid and electric vehicles	9	0.0166667
Mechatronics	7	0.0129630
Noise, vibration and harshness	7	0.0129630
Powertrain	6	0.0111111
Stress and fracture	6	0.0111111
High strength steel	5	0.0092593
Crashworthiness	4	0.0074074
Other	4	0.0074074

- Table 3: Top 5 institutions with the most researchers in the top research field
  - Identify the most dominant institution in a field

```
top_institutions = researchers %>%
  filter(Tag.1 == top_research_field) %>%
  count(Institution, sort = TRUE)%>%
  rename(Reseacher_count = n)%>% head(5)
kable(top_institutions, caption = paste("Top 5 Institutions in", top_research_field))
```

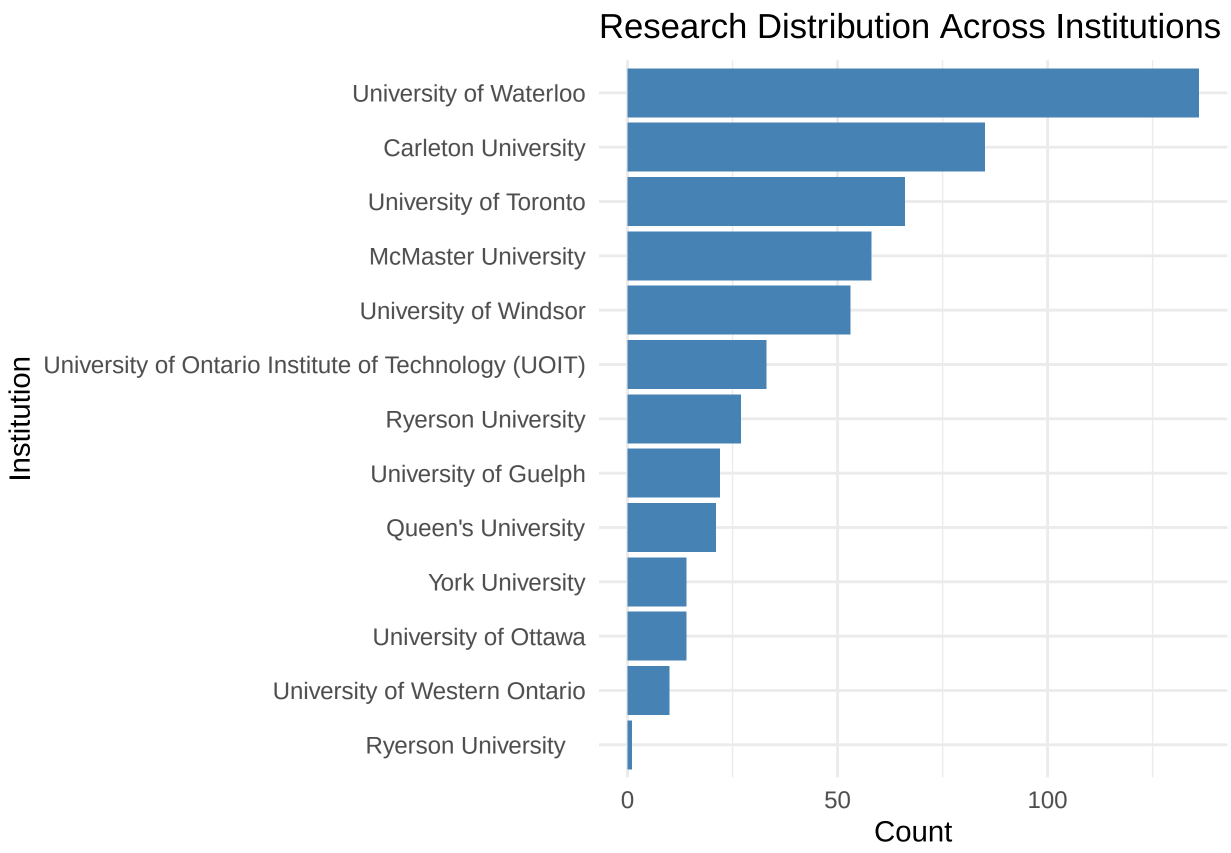
Table 2: Top 5 Institutions in Networks and security

Institution	Reseacher_count
Carleton University	33
University of Waterloo	15
Ryerson University	6
University of Ontario Institute of Technology (UOIT)	3
University of Ottawa	3

Graph & Visualizations

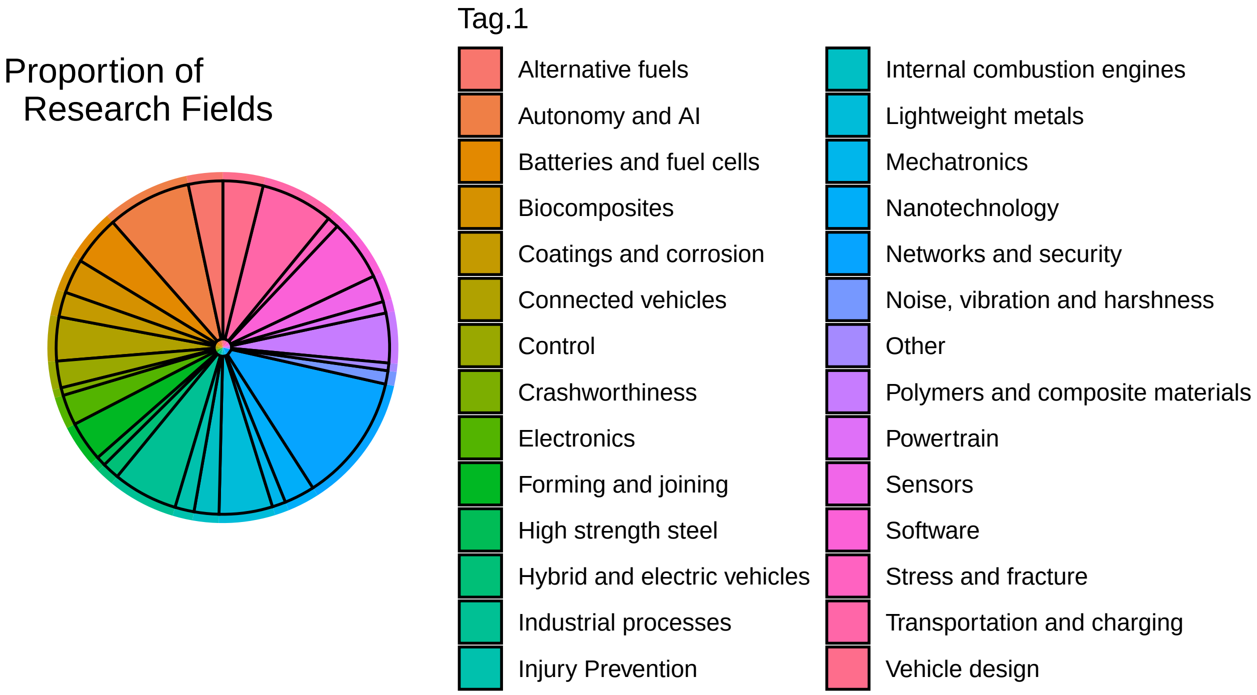
- Bar chart: Research distribution across institution

```
ggplot(researchers %>% count(Institution, sort = TRUE), aes(x = reorder(Institution, n), y = n)) +
 geom_bar(stat = "identity", fill = "steelblue") +
 coord_flip() +
 theme_minimal() +
 labs(title = "Research Distribution Across Institutions",
      x = "Institution", y = "Count")
```



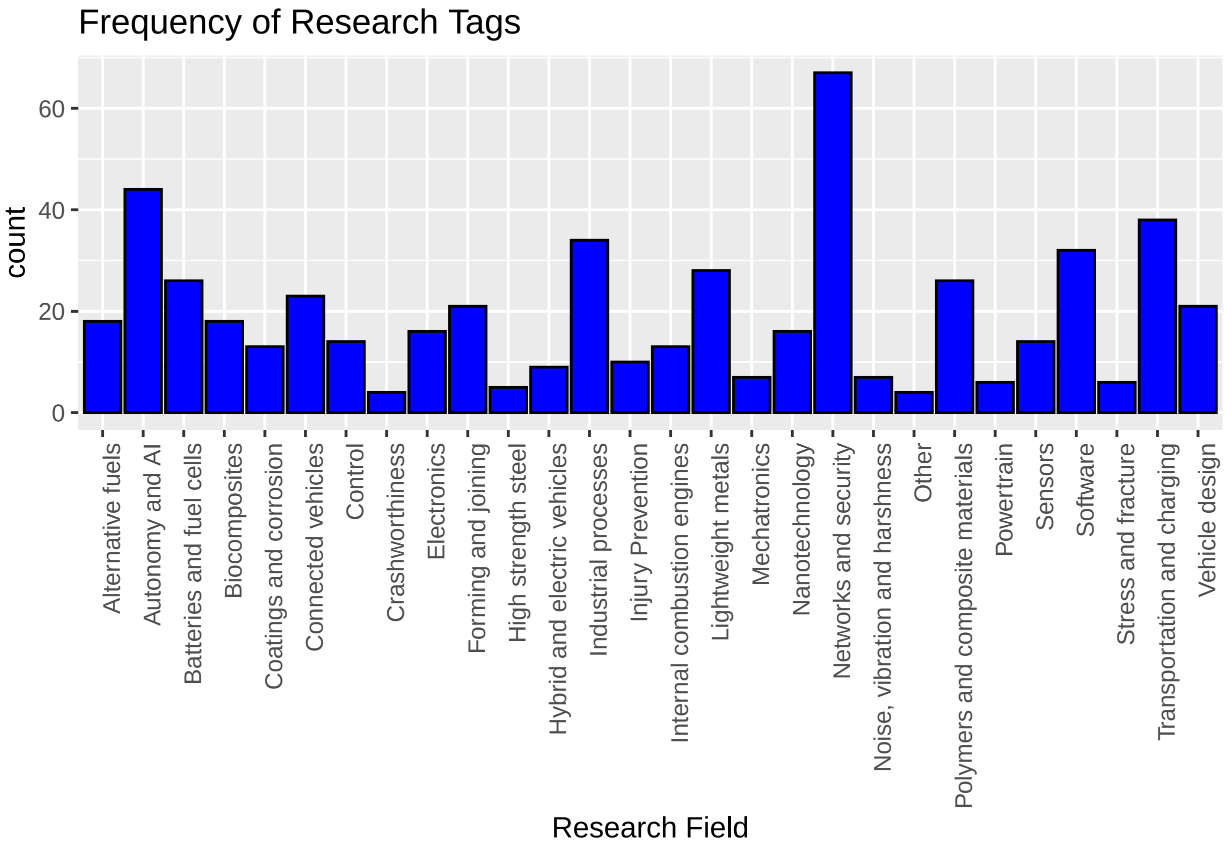
- Pie chart: Proportion of of different research fields

```
ggplot(research_count, aes(x = "", y = Count, fill = Tag.1)) +
 geom_bar(stat = "identity", width = 1) +  geom_col(color = "black") +
 coord_polar("y", start = 0) +
 theme_void() +
 labs(title = "Proportion of
 Research Fields")
```



- History, frequency of research tag

```
ggplot(researchers, aes(x = Tag.1)) +
 geom_bar(fill = "blue", color = "black") +
 theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
 labs(title = "Frequency of Research Tags", x = "Research Field", y = "count")
```



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
- Hypothesis Testing
- Bootstrapping

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)                        -1.823e-01  6.055e-01  -0.301   0.7633
## is_Alternative.Fuels                1.435e+00  8.295e-01   1.730   0.0836 .
## is_Autonomy.and.AI                  1.272e-14  6.770e-01   0.000   1.0000
## is_Batteries.and.Fuel.Cells         6.523e-01  7.274e-01   0.897   0.3699
## is_Biocomposites                    4.055e-01  7.692e-01   0.527   0.5981
## is_Coatings.and.Corrosion           2.817e-02  8.223e-01   0.034   0.9727
## is_Connected.Vehicles               8.109e-01  7.472e-01   1.085   0.2778
## is_Control                         -1.117e+00  8.893e-01  -1.256   0.2091
## is_Crashworthiness                  1.281e+00  1.304e+00   0.982   0.3259
## is_Electronics                      4.336e-01  7.878e-01   0.550   0.5820
## is_Forming.and.Joining              1.629e+00  8.219e-01   1.982   0.0474 *
## is_High.Strength.Steel             -2.231e-01  1.095e+00  -0.204   0.8386
## is_Hybrid.and.Electric.Vehicles    -4.082e-02  9.037e-01  -0.045   0.9640
## is_Industrial.Processes             3.001e-01  6.962e-01   0.431   0.6664
## is_Injury.Prevention                1.030e+00  9.181e-01   1.122   0.2621
## is_Internal.Combustion.Engines      1.386e+00  8.944e-01   1.550   0.1212
## is_Lightweight.Metals               7.701e-01  7.226e-01   1.066   0.2866
## is_Mechatronics                    -1.054e-01  9.747e-01  -0.108   0.9139
## is_Nanotechnology                   6.931e-01  7.958e-01   0.871   0.3838
## is_Networks.and.Security           -5.333e-01  6.590e-01  -0.809   0.4184
## is_Polymers.and.Composite.Materials  4.925e-01  7.240e-01   0.680   0.4964
## is_Powertrain                       1.823e-01  1.017e+00   0.179   0.8577
## is_Sensors                         -4.055e-01  8.233e-01  -0.493   0.6224
## is_Software                        -4.643e-01  7.108e-01  -0.653   0.5136
## is_Stress.and.Fracture              8.755e-01  1.057e+00   0.828   0.4074
## is_Transportation.and.Charging     -3.567e-01  6.926e-01  -0.515   0.6066
## is_Vehicle.Design                   4.700e-01  7.491e-01   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: 748.33  on 539  degrees of freedom
## Residual deviance: 697.27  on 513  degrees of freedom
## AIC: 751.27
##
## 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_Injury.Prevention	1.0296194	2.80
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	2.01
is_Batteries.and.Fuel.Cells	0.6523252	1.93
is_Polymers.and.Composite.Materials	0.4924765	1.64
is_Vehicle.Design	0.4700036	1.61
is_Electronics	0.4336360	1.55
is_Biocomposites	0.4054651	1.51
is_Industrial.Processes	0.3001046	1.36
is_Powertrain	0.1823216	1.21
is_Coatings.and.Corrosion	0.0281709	1.03
is_Autonomy.and.AI	0.0000000	1.01
is_Hybrid.and.Electric.Vehicles	-0.0408220	0.97
is_Mechatronics	-0.1053605	0.91
(Intercept)	-0.1823216	0.84
is_High.Strength.Steel	-0.2231436	0.81
is_Transportation.and.Charging	-0.3566749	0.71
is_Sensors	-0.4054651	0.67
is_Software	-0.4643056	0.63
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: 0.494197 0.581901 0.5850287 0.5702381
## Average AUC score: 0.5578412
```

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