STAA57 Group Project

What Are The Dominant Research Areas in Ontario's Automotive Sector & How do Institutions Specialize In Different Fields

Minh Tran - 1006804914, Swajeet Jadhav - 1009888276, Rayaan Syed - 1010231081

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), automotive research facilities in Ontario must focus their efforts and resources on the right areas. This will ensure a significant contribution to the academic world while assisting the government in implementing effective policies and educational institutions to better educate, prepare the next generation of laborers.

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 the 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 records a researcher's information regarding their workplace and the research areas of expertise. A supporting table providing specific 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"
```

Some of the most important variables that will mainly be used in this report are

1. Institution: Name of the researcher's 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 has one
5. Tag. 1 ~ 5: Categorization of researchers research areas

Data clean up

Based on the data set summary, our team 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.

2. Data Overview Analysis

2.1. Descriptive Statistic

Table 1: Summary statistics of research areas

Research_Area	Count	Frequency
Networks and security	92	0.1703704
Autonomy and AI	78	0.1444444
Transportation and charging	64	0.1185185
Polymers and composite materials	62	0.1148148
Lightweight metals	61	0.1129630
Batteries and fuel cells	59	0.1092593
Industrial processes	57	0.1055556
Forming and joining	56	0.1037037
Connected vehicles	55	0.1018519
Nanotechnology	51	0.0944444
Hybrid and electric vehicles	47	0.0870370
Sensors	47	0.0870370
Software	47	0.0870370
Vehicle design	37	0.0685185
Control	34	0.0629630
Injury Prevention	33	0.0611111
Electronics	32	0.0592593
Alternative fuels	31	0.0574074
Biocomposites	30	0.0555556
Coatings and corrosion	30	0.0555556
Internal combustion engines	28	0.0518519
Stress and fracture	27	0.0500000
Powertrain	26	0.0481481
Mechatronics	16	0.0296296
High strength steel	15	0.0277778
Noise, vibration and harshness	13	0.0240741
Crashworthiness	12	0.0222222
Other	4	0.0074074
Total	540	1.0000000

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 portion of researchers at 17%, 14%, and 12%, respectively.

Please note that one researcher could engage in researches from multiple areas. Therefore, the counting number might contains duplicates and the **Total** row is no the sum of all research field but the total number of researchers in the data set.

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

Institution	Reseacher_count
University of Toronto	6

Table 3: Top 5 Institutions in Autonomy and AI

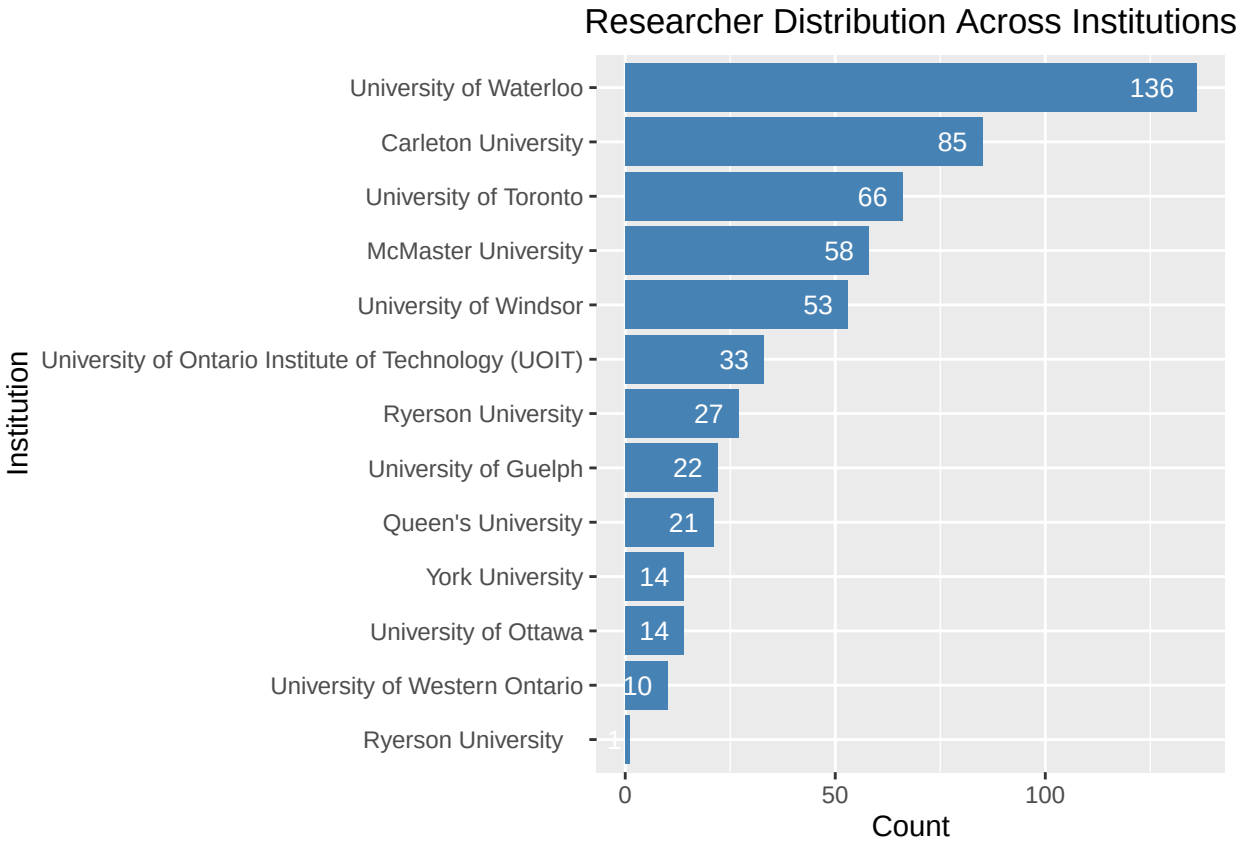
Institution	Reseacher_count
Carleton University	23
University of Waterloo	16
University of Toronto	11
Ryerson University	6
University of Windsor	5

Table 4: Top 5 Institutions in Transportation and charging

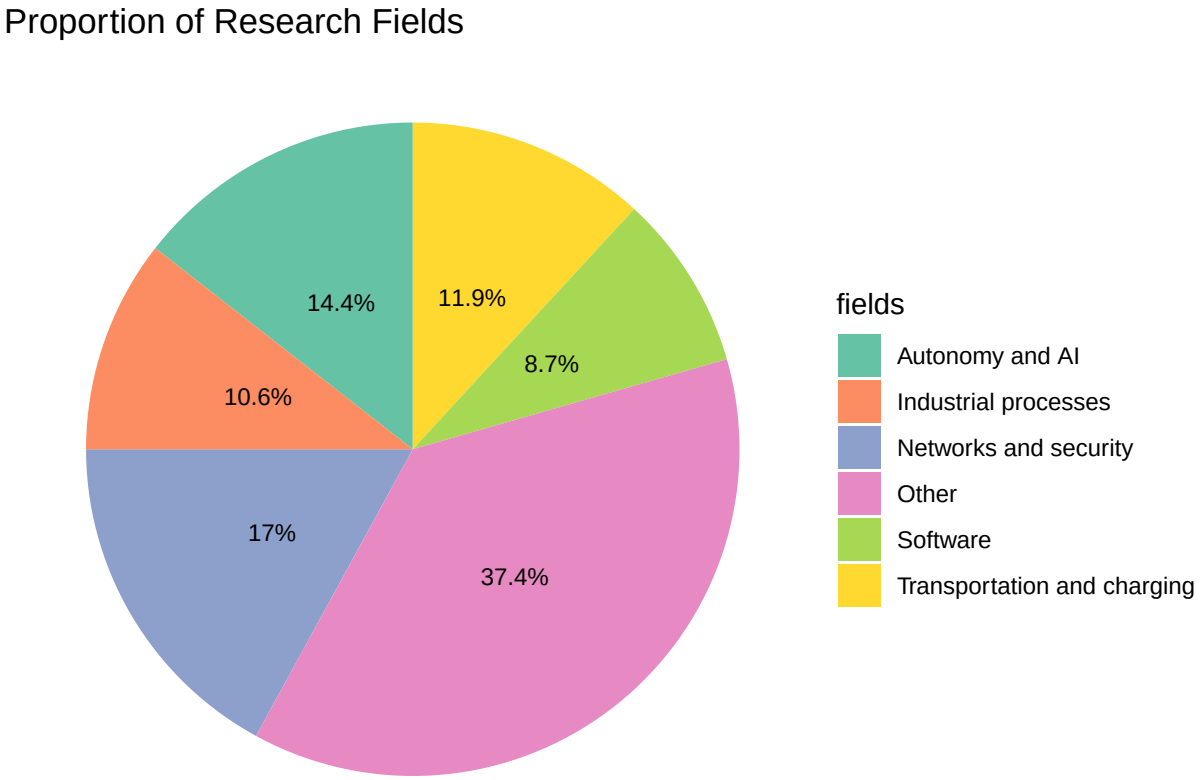
Institution	Reseacher_count
Carleton University	13
University of Waterloo	12
Ryerson University	11
University of Toronto	6
York University	6

Carleton University has a significant number of researchers across three leading research fields at 37, 23, and 13, accordingly. Followed closely are 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 the nation's research agenda

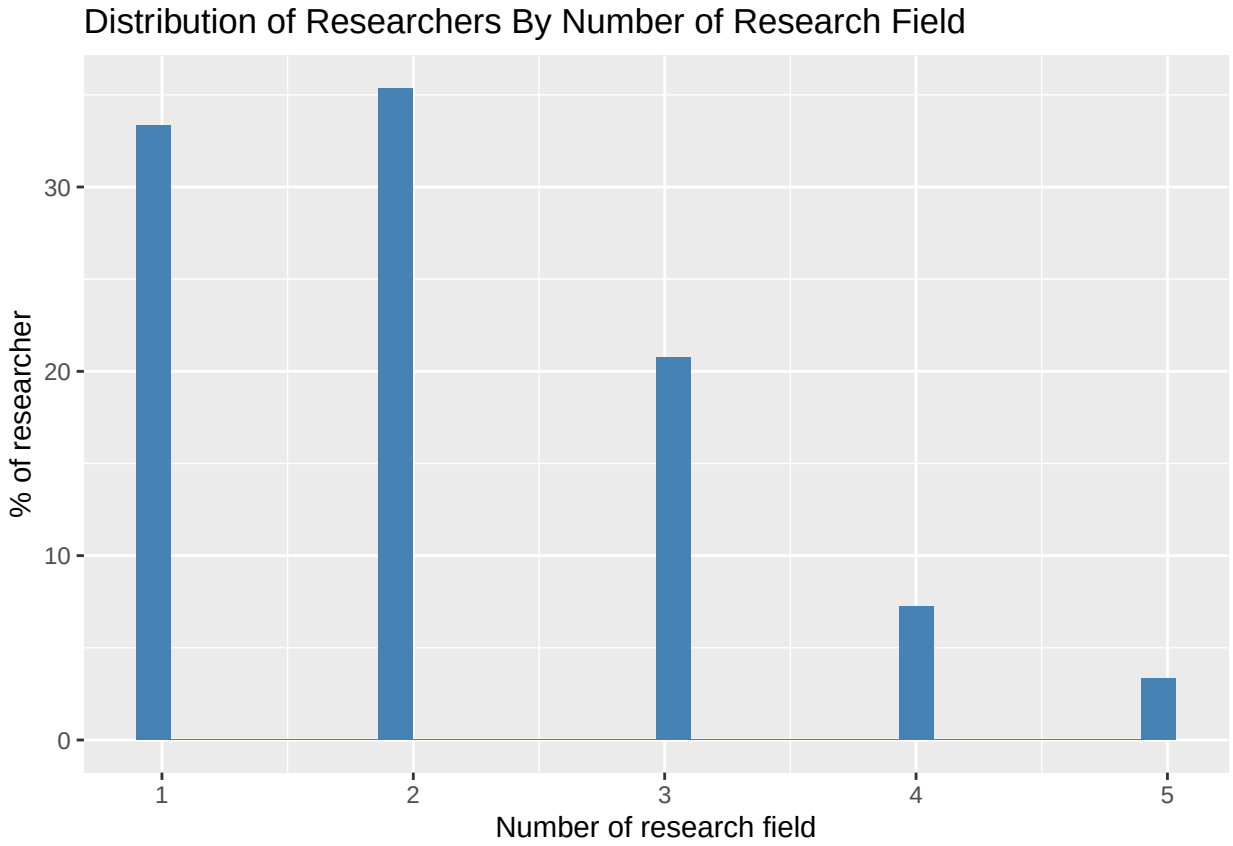
2.2. Graph & Visualizations



The graph presents a visual representation of the total number of researchers employed by each institution. The findings reinforce our conclusions drawn in Section 2.1, confirming that leading institutions such as the University of Waterloo, Carleton University, University of Toronto continue to have the highest number of researchers engaged with the automotive industry at 136, 85, and 66, respectively. Notably, while Carleton University leads in terms of researchers specializing in key fields, the University of Waterloo has the largest number of researchers overall. This data may serve as a valuable resource for policymakers and investors, enabling them to strategically allocate support to institutions that have the potential to contribute significantly to automotive research.



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 the top five research fields account for more than 50% of researchers in the data set showing the dominance of the top 5 group compared to the remaining 23 fields. Another key point from this visualization is that research results from all five fields mostly support the development of Autonomous Electrical vehicles such as Tesla.This is likely due to the rising trend of Artificial Intelligence and replacing fossil fuel vehicles with sustainable, environmentally friendly transportation.

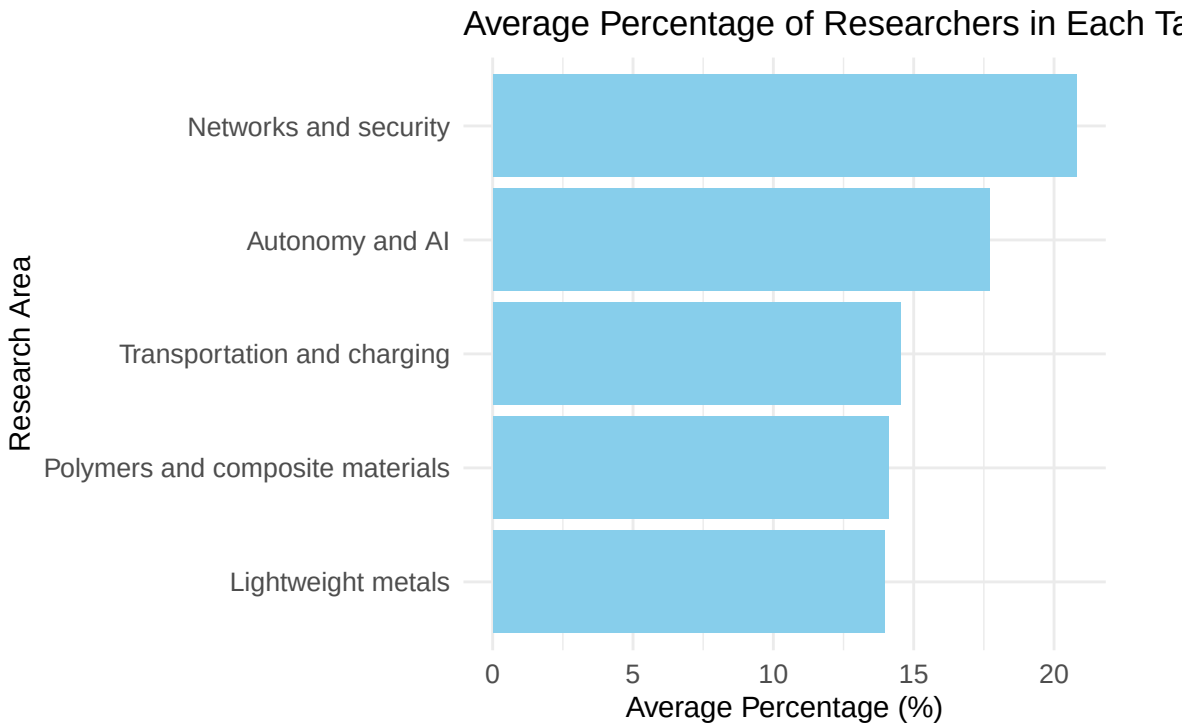


This diagram illustrates the distribution of researchers based on the number of research fields they are associated with, as indicated by the "Tag" columns. If a researcher has a value in only one "Tag" column, they are considered to be specialized in a single field. Similarly, researchers with values 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.

3. Statistical Analysis

3.1. Bootstrapping

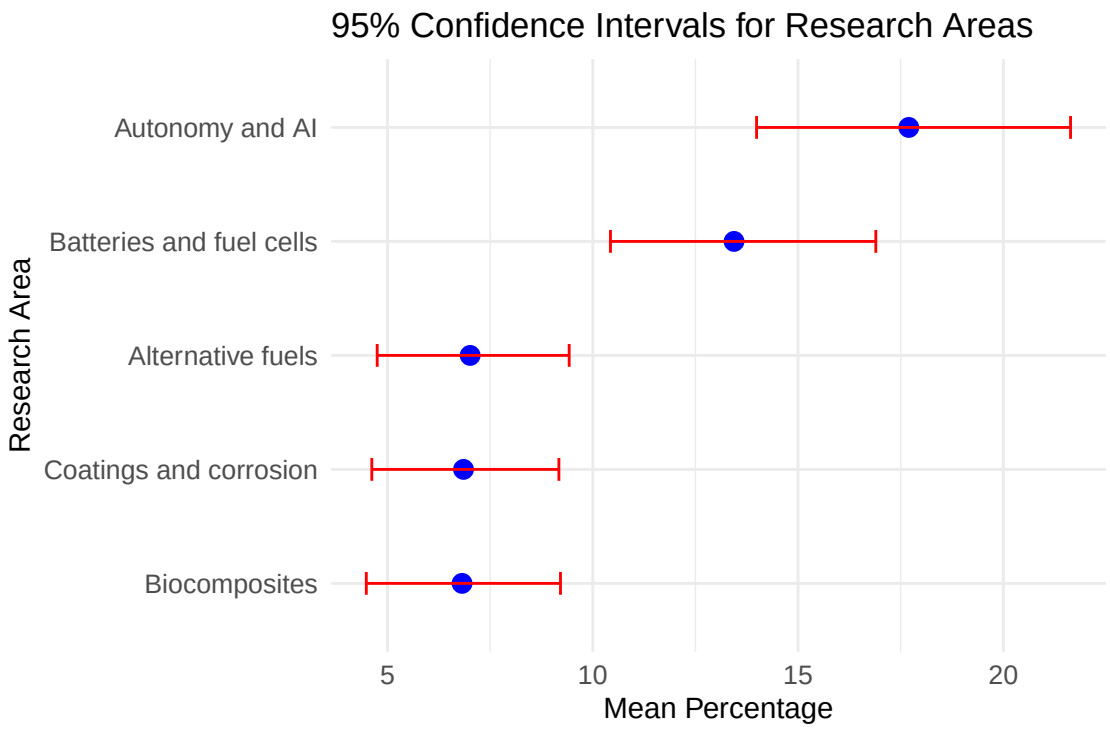
To determine whether the most prominent research area demonstrates a statistically significant lead over others, we employ a bootstrapping approach. Specifically, we aim to assess whether the observed dominance of **Networks and Security** holds consistently across repeated samples and whether this trend is statistically meaningful when compared to the second mot represented field.



3.2. Confidence Intervals

The bootstrapping procedure involved re-sampling the data set 1000 times. In each iteration, the proportion of researchers associated with each research area was calculated relative to the total number of researchers in the original data set. The average researcher percentage for each research field was then computed across all iterations and visualized above. The resulting bar chart illustrates that **Networks and Security** and **Autonomy and AI** exhibit the highest average representation across the simulations.

To quantify the variability and reliability of these estimates, we computed 95% confidence intervals (CIs) for each research area's average researcher percentage, as shown below.



Although **Networks and Security** has the highest mean proportion, the CI for Autonomy and AI partially overlaps with it, suggesting that a direct comparison is is required to confirm whether this difference is statistically significant.

3.3. Hypothesis testing

To formally test the hypothesis that **Networks and Security** leads over **Autonomy and AI**, we perform an independent one-sided t-test comparing their bootstrapped distributions. The hypotheses are defined as follows:

Null Hypothesis (H_0): $\mu_{\textbf{NetworkSecurity}} \leq \mu_{\textbf{AutonomyAI}}$

Alternative Hypothesis (H_1): $\mu_{\textbf{NetworkSecurity}} > \mu_{\textbf{AutonomyAI}}$

```
## Independent t-test: Networks and Security vs. Autonomy and AI
## Mean % Network and Security: 20.80283
## Mean % ( Autonomy and AI ): 17.69679
## T-statistic: 33.33881
## P-value: 2.066354e-194
## Reject H0. Networks and Security is significantly higher than the other tags. (p < 0.05).
```

The bootstrapped mean percentages were 20.8% for **Networks and Security** and 17.7% for **Autonomy and AI**. The resulting t-statistic 33.34 with a p-value of 2.06e-194 indicates that there is less than a 1% chance that we will observe a sample where the average percentage of researchers engaging in **Networks and Security** is less than that of **Autonomy and AI** or something more extreme assuming that is true for the entire population. Therefore, we could safely reject the null hypothesis and confirm that the portion of researchers involved in research in **Networks and Security** is significantly higher than other research areas.

4. Predictive modeling (Regression)

4.1. Logistic Regression Model

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 trends are essential indicators of the areas prioritized by the Canadian government and industry stakeholders in the automotive sector.

To analyze the factors influencing research funding, we employed a logistic regression model with the primary dependent variable being **is_Funded**. This variable takes a value of 1 if the research has secured funding from the Canada Research Chair program or other similar grants, and 0 otherwise. As each researcher could be involved in one or multiple research fields, combining all research fields in the model will create too many variables for the scope of this research. Therefore, the independent variables for our model will focus on the various researchers' primary fields as categorized in **Tag.1** column of the data set.

4.2. Result and Key Findings

Table 5: Top 5 Independent variable with highest odd ratio

	coefficient	odd
is_Forming.and.Joining	1.3217558	3.75
is_Alternative.Fuels	1.1275998	3.09
is_Internal.Combustion.Engines	1.0788097	2.95
is_Crashworthiness	0.9734491	2.65
is_Injury.Prevention	0.7221347	2.06

A substantial proportions of the variables in the 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. Nevertheless, the model provides valuable insights into general trends in funding allocation by the Canadian government and major investors.

Furthermore, certain research fields still stood out in terms of their impact on 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.

4.3. 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: 0.6448804 0.6377515 0.584991 0.5291803
```

```
## Average AUC score: 0.5992008
```

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.

5. Summary

In conclusion, our research results shows that **Networks and Security**, **Autonomy and AI** and **Transportation and Charging** are the leading fields, collectively attracting more than 30% number of researchers in the data set. Among the institutions, Carleton University, University of Toronto and University of Waterloo emerged as key contributors, reinforcing their reputation as hubs for technological research and innovation. Through bootstrapping and hypothesis testing, we confirmed that **Networks and Security** holds a statistically significant lead over other research categories. This insight highlights the growing emphasis on cyber security and connectivity in the development of autonomous and electric vehicle. Additionally, our logistic regression model explored funding trends suggesting that while not all research fields significantly impact funding likelihood, certain areas, such as **Forming and Manufacturing** show stronger associations with receiving financial support. These insights should offer guidance for research institutions, policymakers and investors regarding academic program alignment or resource allocation and optimization. As Ontario continues to make major strive in automotive technology, strategic investment in high-impact research areas will be crucial for driving advancements in AI-powered transportation and sustainable mobility solutions.

6. Appendix

```
# Read the main data set in as researchers and tags file as researcher_tags
researchers = read.csv("ontarioautoresearchers.csv")
## Part 1
#Display data set variables
names(researchers)[1:10]
# Remove outliers
researchers = researchers %>%
  filter(!Institution %in% c("Brock University", "Lakehead University",
                             "Laurentian University", "Royal Military College"))
## Part 2:
# Summary number and percentage of researchers in each research field
research_summary = researchers %>%
  pivot_longer(
    cols = starts_with("Tag."),
    names_to="Tag_Number",
    values_to = "Research_Area") %>%
  filter(!is.na(Research_Area) & Research_Area != "")
research_count = research_summary %>%
  group_by(Research_Area) %>%
  summarise(Count = n()) %>%
  mutate(Frequency = Count / nrow(researchers)) %>%
  arrange(desc(Count))
total_row = tibble(Research_Area = "**Total**", Count = 540, Frequency = 1.00)
research_count = bind_rows(research_count, total_row)
kable(research_count, caption = "Summary statistics of research areas")
# Top 5 institutions with highest number of researchers in each leading fields
## Research field with highest number of researchers
```

```r
top_institutions = research_summary %>%
  filter(Research_Area == research_count$Research_Area[1]) %>%
  count(Institution, sort = TRUE)%>%
  rename(Reseacher_count = n)%>% head(5)
kable(top_institutions, caption = paste("Top 5 Institutions in", research_count$Research_Area[1]))
## Research field with second highest number of researchers
top_institutions = research_summary %>%
  filter(Research_Area == research_count$Research_Area[2]) %>%
  count(Institution, sort = TRUE)%>%
  rename(Reseacher_count = n)%>% head(5)
kable(top_institutions, caption = paste("Top 5 Institutions in", research_count$Research_Area[2]))
## Research field with third highest number of researchers
top_institutions = research_summary %>%
  filter(Research_Area == research_count$Research_Area[3]) %>%
  count(Institution, sort = TRUE)%>%
  rename(Reseacher_count = n)%>% head(5)
kable(top_institutions, caption = paste("Top 5 Institutions in", research_count$Research_Area[3]))
##Bar chart of number of researchers in each institution
ggplot(researchers %>%
         count(Institution, sort = TRUE),
       aes(x = reorder(Institution, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = n), hjust = 1.5, size = 3.5, col="white") +
  coord_flip() +
  theme(plot.title=element_text(hjust=0.5)) +
  labs(title = "Researcher Distribution Across Institutions",
       x = "Institution", y = "Count")
## Pie chart for percentage researcher in each research field
fields = c("Networks and security", "Autonomy and AI",
           "Transportation and charging","Industrial processes", "Software",
           "Other")
NWS_portion = research_count$Frequency[research_count$Research_Area=="Networks
                                        and security"]
AA_portion = research_count$Frequency[research_count$Research_Area==
                                        "Autonomy and AI"]
TC_portion = research_count$Frequency[research_count$Research_Area==
                                        "Transportation and charging"]
IP_poriton = research_count$Frequency[research_count$Research_Area==
                                        "Industrial processes"]
Soft_portion = research_count$Frequency[research_count$Research_Area==
                                          "Software"]
Other_porition = 1 - sum(NWS_portion,AA_portion, TC_portion,
                         IP_poriton, Soft_portion)
frequencies = c(NWS_portion,AA_portion,TC_portion,IP_poriton,
                Soft_portion,Other_porition)
# Create a data frame
df = data.frame(fields, frequencies)
# Create the pie chart
ggplot(df, aes(x = "", y = frequencies, fill = fields)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "Proportion of Research Fields") +
```

```
  scale_fill_manual(values = c("#66c2a5", "#fc8d62", "#8da0cb",
                               "#e78ac3", "#a6d854", "#ffd92f")) +
  geom_text(aes(label = paste0(round(frequencies*100, 1), "%")),
            position = position_stack(vjust = 0.5), size = 3)
# Histogram for distribution of researchers by number of research field
reseach_field_num = researchers %>% rowwise() %>%
  mutate(field_num =
           sum(across(starts_with("Tag."), ~ .x != "" & !is.na(.x))))%>%ungroup()

ggplot(reseach_field_num, aes(x=field_num,
                              y = (..count.. / sum(..count..))*100)) +
   geom_histogram(fill = "steelblue", size = 10) +
   labs(title = "Distribution of Researchers By Number of Research Field",
        x = "Number of research field", y = "% of researcher")
# Part 3:
# Bootstrapping
set.seed(50)
research_summary = researchers %>%
  pivot_longer(
    cols = starts_with("Tag."),
    names_to = "Tag_Number",
    values_to = "Research_Area") %>%
  # Sanity Check
  filter(!is.na(Research_Area) & Research_Area != "")
num_samples = 1000  # Number of bootstrap samples
bootstrap_results = replicate(
  # Replicate data sampling and summarization 1000 times
  num_samples,
  {
    sampled_data = research_summary %>% sample_n(n(), replace = TRUE)

    total_researchers = n_distinct(sampled_data$Researcher.Name)

    sampled_data %>%
      group_by(Research_Area) %>%
      # Percentage of each tag
      summarise(Percentage = n() / total_researchers * 100)
  }, simplify = FALSE
)
boot_results_df = bind_rows(bootstrap_results, .id = "Iteration")
average_percentage_df = boot_results_df %>%
  group_by(Research_Area) %>%
  summarise(Mean_Percentage = mean(Percentage)) %>%
  # Sort by average percentage in descending order for readability
  arrange(desc(Mean_Percentage)) %>% head(5)
# Plot the horizontal graph
ggplot(average_percentage_df, aes(x = reorder(Research_Area, Mean_Percentage),
                                  y = Mean_Percentage)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +  # Flip to make it horizontal
  labs(title = "Average Percentage of Researchers in Each Tag",
       x = "Research Area", y = "Average Percentage (%)") +
  theme_minimal() +
```

```r
  theme(plot.margin = margin(1, 1, 1, 1, "cm"),  # Adjust margins
        axis.text.y = element_text(size = 10),  # Reduce y-axis text size
        axis.text.x = element_text(size = 10))  # Reduce x-axis text size
# Finding and plot 95% Confidence Interval of Each Research Area
ci_table <- boot_results_df %>%
  group_by(Research_Area) %>%
  summarise(
    Mean_Percentage = mean(Percentage, na.rm = TRUE),
    Lower_CI = quantile(Percentage, 0.025, na.rm = TRUE),
    Upper_CI = quantile(Percentage, 0.975, na.rm = TRUE)
  ) %>% head(5)
ggplot(ci_table, aes(x = reorder(Research_Area, Mean_Percentage),
                     y = Mean_Percentage)) +
  geom_point(color = "blue", size = 3) +  # Mean values as points
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI), width = 0.2, color = "red") +
  coord_flip() +
  labs(
    title = "95% Confidence Intervals for Research Areas",
    x = "Research Area",
    y = "Mean Percentage"
  ) + theme_minimal() + theme(plot.margin = margin(1, 1, 1, 1, "cm"),
        axis.text.y = element_text(size = 10),  # Reduce y-axis text size
        axis.text.x = element_text(size = 10))  # Reduce x-axis text size
# T.test: highest and second highest field in terms of researchers percentage
# 2nd-ranked tag
second_place_tag = average_percentage_df$Research_Area[2]
# Prepare data for t-test
network_data = boot_results_df %>%
  filter(Research_Area == "Networks and security") %>%
  pull(Percentage)
second_place_data = boot_results_df %>%
  filter(Research_Area == second_place_tag) %>%
  pull(Percentage)
# t-test to show Networks and Security > 2nd place
t_test_result = t.test(
  network_data,
  second_place_data,
  alternative = "greater"
)
# Display results
cat("Independent t-test: Networks and Security vs.", second_place_tag, "\n")
cat("Mean % Network and Security:", mean(network_data), "\n")
cat("Mean % (", second_place_tag, "):", mean(second_place_data), "\n")
cat("T-statistic:", t_test_result$statistic, "\n")
cat("P-value:", t_test_result$p.value, "\n") # Print results
if (t_test_result$p.value < 0.05) {
  cat("Reject H0. Networks and Security is
      significantly higher than the other tags. (p < 0.05).")
} else {
  cat("No significant difference (p >= 0.05).") # Check hypothesis
}
# Part 4: Regression analysis
#Prepare data for regression model
```

```r
reg_data = na.omit(researchers)
#This code dynamically create a dummy variable for all research field columns
# Dynamically get all research field indicating columns after "Tag.5 column"
start_col = which(names(reg_data) == "Tag.5") + 1
research_indicator_col = names(reg_data)[start_col:length(names(reg_data))]
# Add new summary variables for each of the research field indicating columns
for (col in research_indicator_col){
  new_col_name = paste0("is_", col)
  reg_data[[new_col_name]] = ifelse(tolower(reg_data[["Tag.1"]])==
                                      tolower(gsub("\\."," ", col)),1,0)
}
reg_data=reg_data%>%
  mutate(is_Funded = ifelse(Research.Chairs.Grant.Funding!="",1,0))
reg_data = reg_data %>% select(contains("is_"))
# Logistic regression model
formula =
  as.formula(paste("is_Funded ~", paste(setdiff(names(reg_data), c("is_Funded", "is_Vehicle.Design")),
regression = glm( formula , family = binomial, data = reg_data)
summary (regression)
reg_result = data.frame(coefficient = coef(regression),
                        odd = ceiling(exp(coef(regression)) * 10^2)/10^2)
reg_result = reg_result %>% arrange(desc(coefficient)) %>% head(5)
kable(reg_result, caption = "Top 5 Independent variable with highest
      odd ratio")
# Validation using k-fold method
k = 4
reg_validation = reg_data %>%
  mutate(group_ind = sample(c(1:k), size=nrow(reg_data),replace = T))
c.index = vector()
for (i in 1:k){
  reg_data.train = reg_validation %>% filter(group_ind != i)
  reg_data.test = reg_validation %>% filter(group_ind == i)
  logit.mod = glm(is_Funded ~ ., family = binomial, data = reg_data.train)
  pi_hat = predict(logit.mod, newdata=reg_data.test, type="response")
  m.roc=roc(reg_data.test$is_Funded ~ pi_hat)
  c.index[i] = auc(m.roc)
}
cat("AUC score:",c.index )
cat("Average AUC score:", mean(c.index))
```