Diverse Data Hub

Siddarth Subrahmanian, Francisco Ramírez, Azin Piran

2025-05-05

Table of contents

Introduction and Background

As data becomes central to shaping decisions focused on societal issues, it is critical that data science and statistics students engage not only with technical concepts but also with the ethical and social contexts of the data they analyze.

However, many instructors face challenges in incorporating Equity, Diversity, and Inclusion (EDI) related themes into their teaching due to a lack of accessible, curated datasets and resources.

This project aims to address that need by developing a user-friendly, openly available resource that empowers educators to bring EDI topics into their lecture rooms.

This project is being developed in collaboration with Katie Burak PhD, Assistant Professor of Teaching in the Department of Statistics at the University of British Columbia.

Partner Needs and Project Motivation

The motivation behind this project comes from a growing need for educational resources that enable engagement with EDI topics through data-driven learning. This initiative seeks to fill that gap by providing two key deliverables:

- An R data package containing selected datasets focused on EDI-related topics, such as gender equity, socio-economic inclusion, and more. These datasets, provided by the partner, will be curated and documented for use in educational settings, adhering to the FAIR principles: Findability, Accessibility, Interoperability, and Reusability.

The FAIR data principles are foundational for data stewardship and reuse (GO FAIR Initiative 2024).

- A Quarto-based, GitHub-Pages-deployed, educational website featuring simple, interpretable teaching examples and data science projects. These examples will guide instructors through exploratory, inferential, and predictive analyses while offering contextual narratives that highlight real-world relevance and social impact.

By addressing these needs, this project not only tackles the technical challenges of data accessibility but also aims to foster data literacy through content that reflects real-world social issues.

Proposed Deliverables

Our project is designed with the goal of long-term utility, adaptability, and extensibility in mind. To meet these goals, we structured our deliverables around three key principles:

- **Modularity** Each component of the project—datasets, notebooks, and the website can function independently. For example, educators may choose to use only the R package in their own materials, or they might rely solely on the website for teaching demonstrations. This modular structure ensures flexibility for different user needs and teaching styles.
- **Reusability** All tools and content are developed with reuse in mind. The notebook template can be adapted for new datasets, and the R package can be extended to include additional topics in the future. Each dataset includes documentation and metadata that allows instructors and students to reuse the data across multiple educational contexts without additional preparation.
- **Scalability** The project is built to grow. New datasets can be added to the package and linked through the website with minimal extra work, thanks to a standardized structure. Each deliverable is designed to scale across:
- Additional datasets
- New exploratory questions and case studies
- Broader institutional or cross-course adoption

By designing our outputs to meet these three core principles, we ensure that the final product is not only impactful at launch but also sustainable and expandable over time.

a. R Data Package

We will package the selected datasets into a standalone R package, following the standards and best practices commonly used in well-established R data packages. Each dataset will include:

- Cleaned and standardized data
- Metadata and source attribution
- A detailed description of the dataset's background and relevance to EDI topics
- Suggestions for discussion questions or learning activities

This format allows educators to easily install and access datasets in a consistent format across institutions.

b. Notebook template

We are creating a reusable notebook template written in Markdown to ensure consistency across exercises.

View the Template on GitHub

This will be used as a base format for each dataset's analysis and made available on the website and GitHub. The sections include:

- **About the data**: Source, description, relevance
- **Case Study Objective**: Framing a meaningful EDI-related question
- **Data Wrangling**: Cleaning steps and justifications
- **Exploratory Data Analysis**: Descriptive plots, summaries
- **Statistical Modeling or Inference**: (if applicable)
- **Discussion and Interpretation**: Real-world implications

We are using this template to make sure that all the key steps are followed in every analysis we conduct. This consistency allows us to maintain quality and structure across all datasets. It also helps users of the website easily compare the examples, understand how different datasets were explored, and apply similar approaches to their own data questions.

See Appendix A for a full example of the template structure.

c. Quarto Website

The strategy for the website's architecture focuses on creating an intuitive and scalable structure to showcase the project's datasets, and example analysis notebooks. The website will be organized to ensure that users can easily access important sections, including dataset descriptions, analysis notebooks, and downloadable resources. As shown in the figure (Figure 1), the layout emphasizes clear navigation paths and consistent design elements across pages to support a smooth user experience.

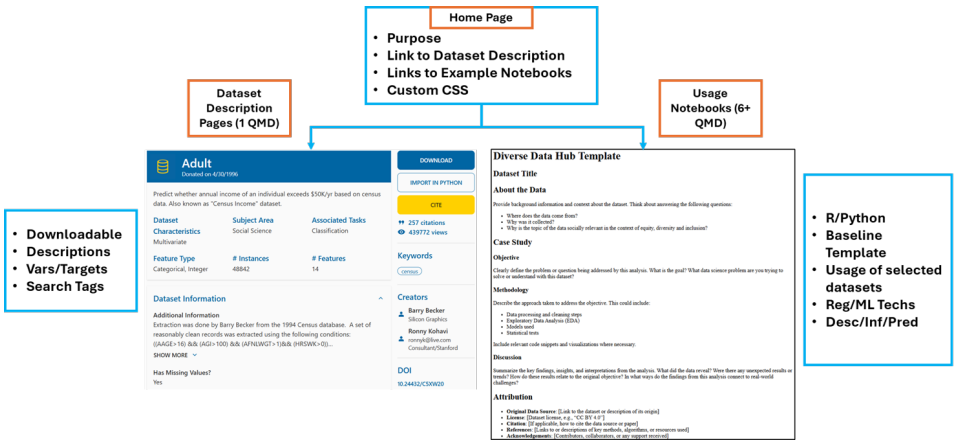


Figure 1: Website Development Strategy

Data Science Approach

This project's strategy is organized into three branches, each focusing on an aspect of the project's deliverables: R Package Development, Analysis, and Website Development. Each branch is designed to ensure the creation of reproducible analysis, and the publication of results in an accessible, user-friendly format. The three branches work in parallel to contribute to the success of the project (Figure 2).

1. R Package Development strategy

   The principles of building R packages are well-documented in (Wickham and Bryan 2023).The steps involve:

- Defining the Purpose and Scope :

  The purpose of the R package is to provide educators with easy access to well-documented, curated datasets focused on equity, diversity, and inclusion (EDI) topics for use in teaching data science and statistics.

  Target Audience: Students, Researchers, Academia

  Scope: The scope of this project is to develop an R package featuring 6 curated EDI-focused datasets with educational context and example analyses for classroom use.

- Set Up the Development Environment

  Software/Tools used:

  – IDE: RStudio

  – Libraries: devtools and usethis packages for development

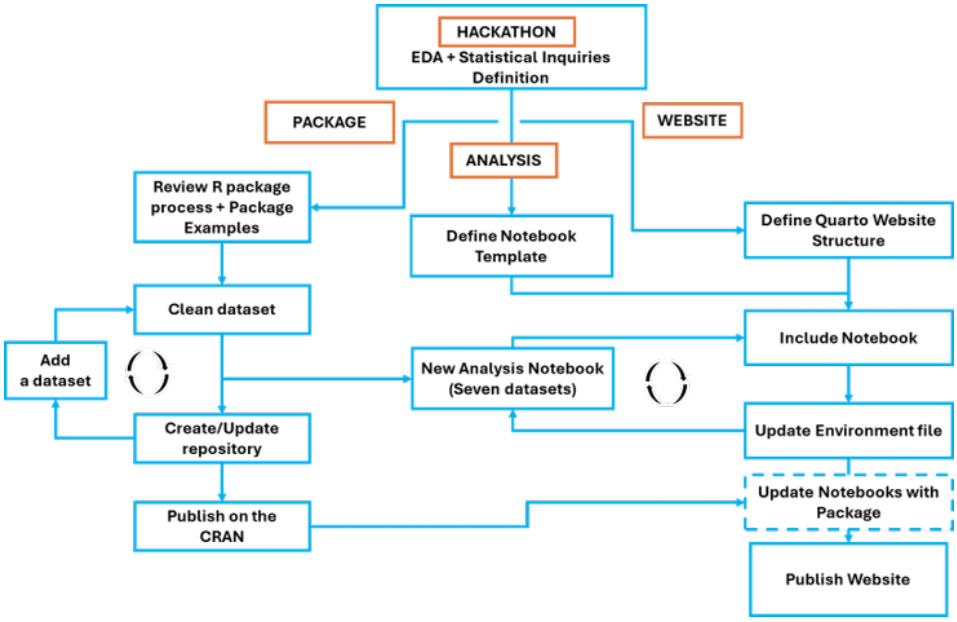


Figure 2: Development Strategy

  - roxygen2 for documentation
  - testthat for unit testing /test attributes of data

- Build Core Package Structure:

  Adding the cleaned dataset into the pkg . as .rda files.

  Different methods proposed:

  - data/: Used for datasets meant to be accessed by users
  - R/sysdata.rda: Used for internal data needed only by package functions.
  - inst/extdata/: Used for raw, non-R-specific files (e.g., CSVs, Excel) accessible to users.

- Testing: Set up testing framework using testthat library. Write unit tests to ensure correctness and stability.
- Add Metadata and Licensing:
  - Add DESCRIPTION file with package name, version, authors, title, and dependencies.
  - License: Use MIT licence
  - Add README.Rmd which contains description about the Packages and usage.

- Check Package Health:
  - Run full checks using: `devtools::check()`
  - Fix any warnings or errors.
- Publish:

  Option A: Publish on GitHub
  - Push to the public GitHub repo.
  - Add installation instructions using devtools

  Option B: Submit to CRAN
  - Ensure the package passes `R CMD` check on multiple platforms.
  - Submit using: `devtools::submit_cran()`
  - Respond to CRAN reviewers' feedback and fix issues if any.
- Maintenance and Updates: Track issues and feature requests (GitHub Issues). Version updates

2. Analysis

The analysis branch is focused on generating reproducible analyses from the datasets. The main activities in this branch include:

- Defining a template for analysis notebooks in Quarto.
- Developing analysis notebooks as datasets become available from the R package.
- Applying exploratory, inferential and predictive analysis techniques using regression, machine learning, and other statistical methods.

3. Website Development

The website branch focuses on developing a user-friendly platform to present the project's work and results. As the interface for the project, the website will offer an interactive space where users can download datasets, and view analysis example notebooks. This branch includes:

- Building the website structure and defining its architecture.
- Integrating the datasets, and analysis notebooks into the website.
- Publishing the website using GitHub Pages, making it accessible to a wider audience.

Project Timeline

Sprint	Dates	Milestone
Planning	April 28 - 30	• Project Kickoff - Hackathon • Project planning
1	May 5 - 9	• R packaging • Data Cleanup • Analysis & creation of template • Website development using Quarto ** Above tasks are for 1 dataset
2	May 12 - 13	• R packaging • Data Cleanup • Analysis & creation of template • Website development using Quarto ** Above tasks are for rest of 5 datasets
3	May 26 - 30	• Integration • Data Analysis - Continuous Improvement
4	June 2 - 6	• Testing • Data Analysis - Continuous Improvement
5	June 9 - 13	• Code Freeze • Release (Prod env)
6	June 16 - 26	• Bug Fixes • Feedback Implementation • Enhancements Stabilization

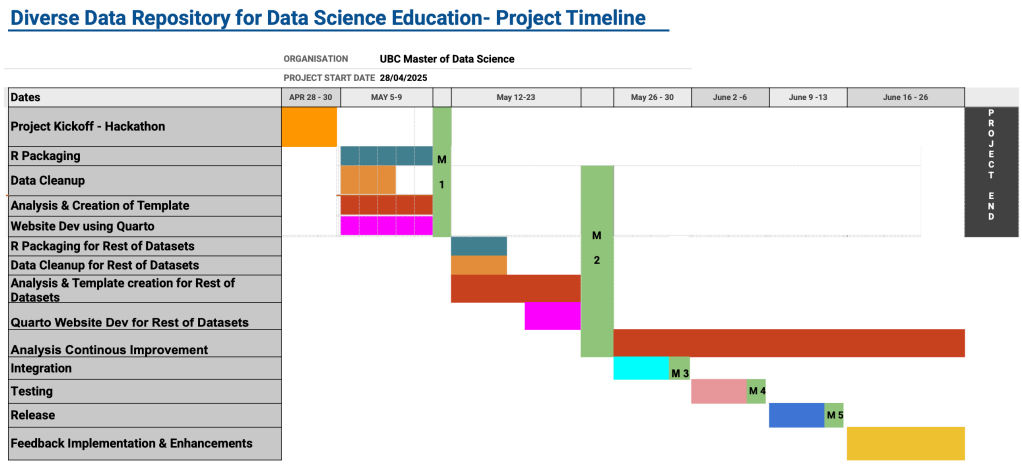


Figure 3: Project Timeline - Gantt Chart

Expected Impact

This project is expected to have a meaningful impact on data education enabling the incorporation of Equity, Diversity, and Inclusion (EDI) topics into the lecture room.

By providing accessible datasets and ready-to-use analysis materials, the project equips educators with practical tools to promote socially relevant learning.

The R package and educational website will support accessible, reproducible, data science instruction, helping students engage not only with technical skills but also with the societal implications of data.

Ultimately, this initiative contributes to a more inclusive and informed data-literate data science community.

Feedback and Iteration

We have actively incorporated inputs from Professors Katie Burak and Ilya, as well as feedback within the team. Early discussions helped narrow the scope, ensuring the project is both feasible and impactful. We will continue to refine our approach based on regular check-ins and informal feedback loops.

Appendices

Appendix A: Quarto Notebook Template

We developed a reusable Quarto (`.qmd`) notebook template that structures every dataset analysis consistently. Below is a sample outline of the template:

Wildfire Dataset

About the Data

This dataset contains information on wildfires in Canada, compiled from official government sources. It includes key variables such as:

- **Fire size (in hectares)**
- **Cause of fire (e.g., lightning, human activity)**
- **Detection method**
- **Response team size**
- **Latitude/longitude of the fire**
- **Weather conditions at the time of fire**

The data was collected to monitor, assess, and respond to wildfire risks across regions. Wildfires have significant environmental, social, and economic impacts—especially for **remote, Indigenous, and underserved communities** that may lack the infrastructure to respond effectively.

From an equity and inclusion perspective, studying wildfire data can help identify **geographic and resource disparities** in fire detection and containment efforts, as well as the disproportionate risks certain populations face due to **climate change** and **infrastructure gaps**.

Case Study

Objective

Can we identify the environmental and human factors most associated with large wildfires (>10 hectares)?

The goal is to explore potential predictors of fire size, such as weather, fire cause, and detection method, and provide insights that could inform early interventions and resource planning.

Methodology

1. Data Cleaning & Processing

- Converted fire size to numeric
- Created a binary variable `large_fire` (TRUE if >10 ha)
- Filtered out incomplete records

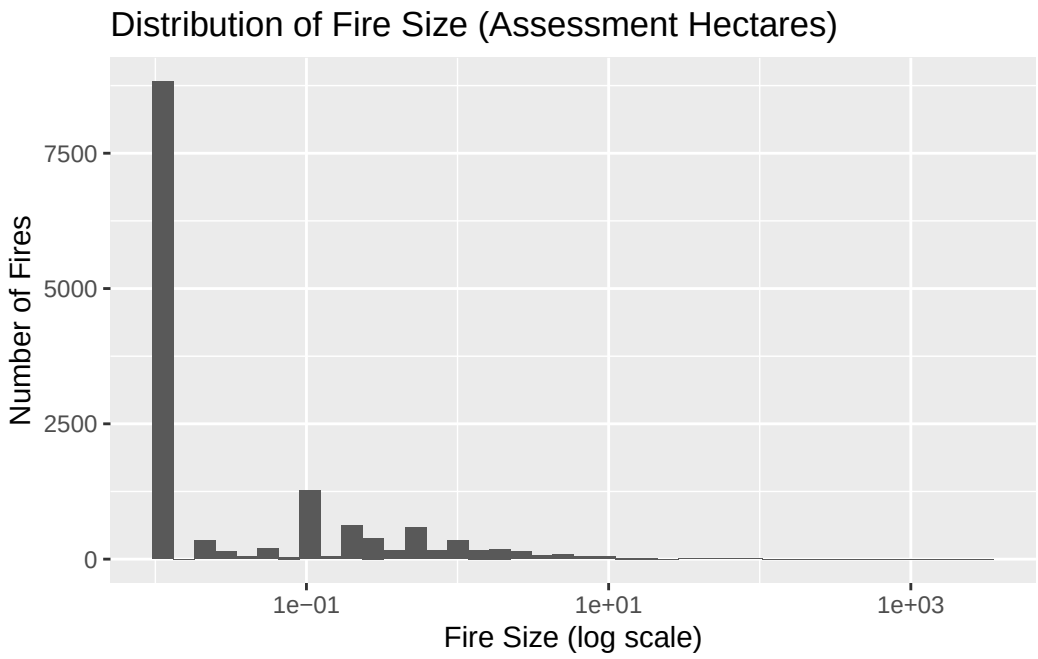
```
library(readr)
library(dplyr)
library(lubridate)
library(gt)

## Reading Data
wildfire_data <- read_csv("data/wildfire.csv")

## Clean and prepare base data
wildfire_clean <- wildfire_data %>%
  filter(!is.na(ASSESSMENT_HECTARES), ASSESSMENT_HECTARES > 0) %>%
  mutate(
    large_fire = ASSESSMENT_HECTARES > 10,
    TRUE_CAUSE = as.factor(TRUE_CAUSE),
    DETECTION_AGENT_TYPE = as.factor(DETECTION_AGENT_TYPE),
    TEMPERATURE = as.numeric(TEMPERATURE),
    WIND_SPEED = as.numeric(WIND_SPEED)
  )

## Drop unused levels for modeling
wildfire_clean <- wildfire_clean %>%
  filter(!is.na(TRUE_CAUSE), !is.na(DETECTION_AGENT_TYPE)) %>%
  mutate(
    TRUE_CAUSE = droplevels(TRUE_CAUSE),
    DETECTION_AGENT_TYPE = droplevels(DETECTION_AGENT_TYPE)
  )
```

2. Exploratory Data Analysis (EDA)

Fire Size Distribution

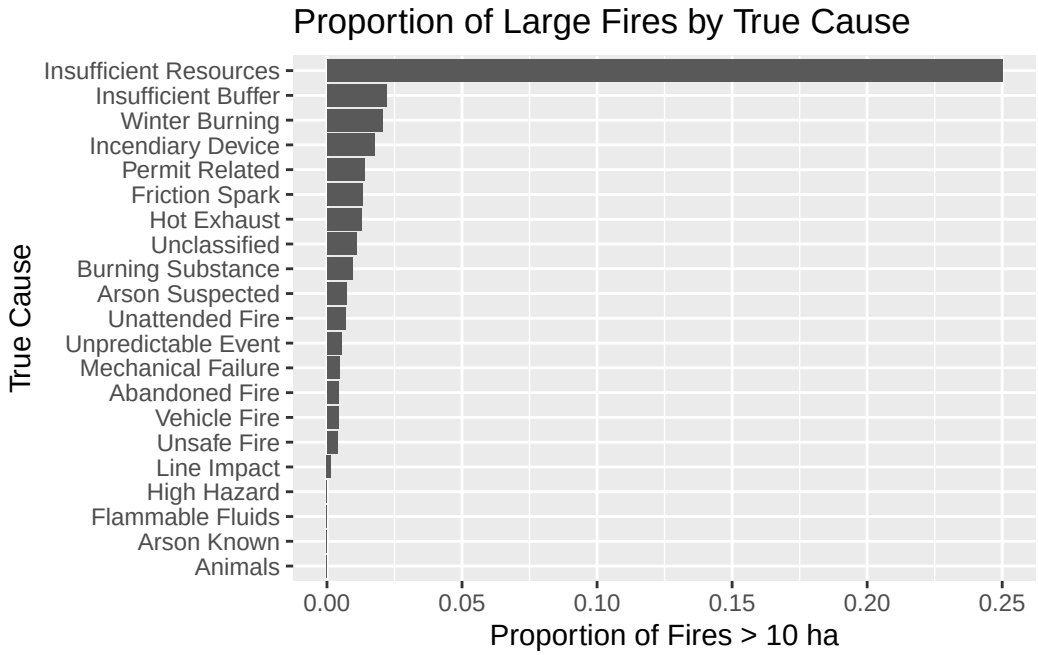
```
library(ggplot2)

ggplot(wildfire_clean, aes(x = ASSESSMENT_HECTARES)) +
  geom_histogram(bins = 40) +
```

```
  scale_x_log10() +
  labs(
    title = "Distribution of Fire Size (Assessment Hectares)",
    x = "Fire Size (log scale)",
    y = "Number of Fires"
  )
```



Proportion of Large Fires by Cause

```
wildfire_clean %>%
  group_by(TRUE_CAUSE) %>%
  summarize(prop_large = mean(large_fire, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(TRUE_CAUSE, prop_large), y = prop_large)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Proportion of Large Fires by True Cause",
    x = "True Cause",
    y = "Proportion of Fires > 10 ha"
  )
```



3. Logistic Regression Model

We build a logistic regression model to predict the likelihood of a fire becoming large based on **temperature**, **wind speed**, and **cause**.

```
library(broom)

model <- glm(
  large_fire ~ TEMPERATURE + WIND_SPEED + TRUE_CAUSE + DETECTION_AGENT_TYPE,
  data = wildfire_clean,
  family = "binomial"
)


# Tidy and clean model output
tidy_model <- broom::tidy(model) %>%
  dplyr::mutate(
    estimate = round(estimate, 3),
    std.error = round(std.error, 3),
    statistic = round(statistic, 2),
    p.value = round(p.value, 4)
  )

# Create a nice table
```

```r
gt_table <- tidy_model %>%
  gt::gt() %>%
  gt::tab_header(
    title = "Logistic Regression Results",
    subtitle = "Predicting Large Fires (> 10 ha)"
  ) %>%
  gt::cols_label(
    term = "Variable",
    estimate = "Estimate (Log-Odds)",
    std.error = "Std. Error",
    statistic = "z value",
    p.value = "p-value"
  ) %>%
  gt::fmt_missing(everything(), missing_text = "-") %>%
  gt::tab_options(
    table.font.size = "small",
    data_row.padding = gt::px(4),
    heading.title.font.size = 16,
    heading.subtitle.font.size = 12
  )

gt_table
```

Discussion

The logistic regression model revealed that **higher wind speeds** are strongly associated with an increased likelihood of a fire becoming large (over 10 hectares), consistent with our expectations about fire spread dynamics.

Surprisingly, **temperature showed a small negative association** with fire size, though this may be influenced by interactions with other environmental factors like humidity or fuel type.

Among causes, **"Insufficient Resources"** and **"Line Impact"** were associated with significantly higher odds of large fires. This suggests that both human-related limitations and infrastructure vulnerability (like power lines) play a role in fire escalation.

The detection agent type showed weak evidence that **fires detected by UNP agents** may be less likely to become large, compared to FPD Staff, but the effect was not statistically strong ($p = 0.09$). Further exploration is needed here, especially considering the early intervention ability of different detection teams.

Logistic Regression Results

Predicting Large Fires (> 10 ha)

Variable	Estimate (Log-Odds)	Std. Error	z value	p-value
(Intercept)	-4.262	0.512	-8.32	0.0000
TEMPERATURE	-0.040	0.012	-3.22	0.0013
WIND_SPEED	0.051	0.007	7.01	0.0000
TRUE_CAUSEAnimals	-15.465	2121.042	-0.01	0.9942
TRUE_CAUSEArson Known	-16.045	2200.700	-0.01	0.9942
TRUE_CAUSEArson Suspected	-0.219	0.671	-0.33	0.7443
TRUE_CAUSEBurning Substance	-0.016	0.450	-0.04	0.9716
TRUE_CAUSEFlammable Fluids	-16.016	4207.738	0.00	0.9970
TRUE_CAUSEFriction Spark	0.401	0.678	0.59	0.5541
TRUE_CAUSEHigh Hazard	-16.284	1926.733	-0.01	0.9933
TRUE_CAUSEHot Exhaust	0.346	0.793	0.44	0.6628
TRUE_CAUSEIncendiary Device	0.489	1.075	0.45	0.6492
TRUE_CAUSEInsufficient Buffer	0.231	0.514	0.45	0.6531
TRUE_CAUSEInsufficient Resources	3.811	1.241	3.07	0.0021
TRUE_CAUSELine Impact	-2.123	1.069	-1.99	0.0470
TRUE_CAUSEMechanical Failure	-0.532	0.790	-0.67	0.5011
TRUE_CAUSEPermit Related	0.326	0.425	0.77	0.4439
TRUE_CAUSEUnattended Fire	0.041	1.064	0.04	0.9696
TRUE_CAUSEUnclassified	0.375	0.791	0.47	0.6355
TRUE_CAUSEUnpredictable Event	-0.738	0.800	-0.92	0.3562
TRUE_CAUSEUnsafe Fire	-0.146	0.401	-0.36	0.7161
TRUE_CAUSEVehicle Fire	-0.631	1.060	-0.60	0.5516
TRUE_CAUSEWinter Burning	0.433	0.454	0.95	0.3409
DETECTION_AGENT_TYPEGRP	-16.143	438.705	-0.04	0.9706
DETECTION_AGENT_TYPELKT	0.023	0.371	0.06	0.9502
DETECTION_AGENT_TYPEUNP	-0.613	0.363	-1.69	0.0907

These findings provide insights into key environmental and operational factors influencing wildfire severity. Importantly, they point to the **need for targeted mitigation strategies** in areas with poor detection access or high infrastructure risks.

In the broader context of equity, this analysis reinforces that **resource constraints and delayed detection**—often more common in remote or underfunded regions—can amplify wildfire impacts. Data-informed strategies can help ensure **more equitable protection** against climate-driven disasters.

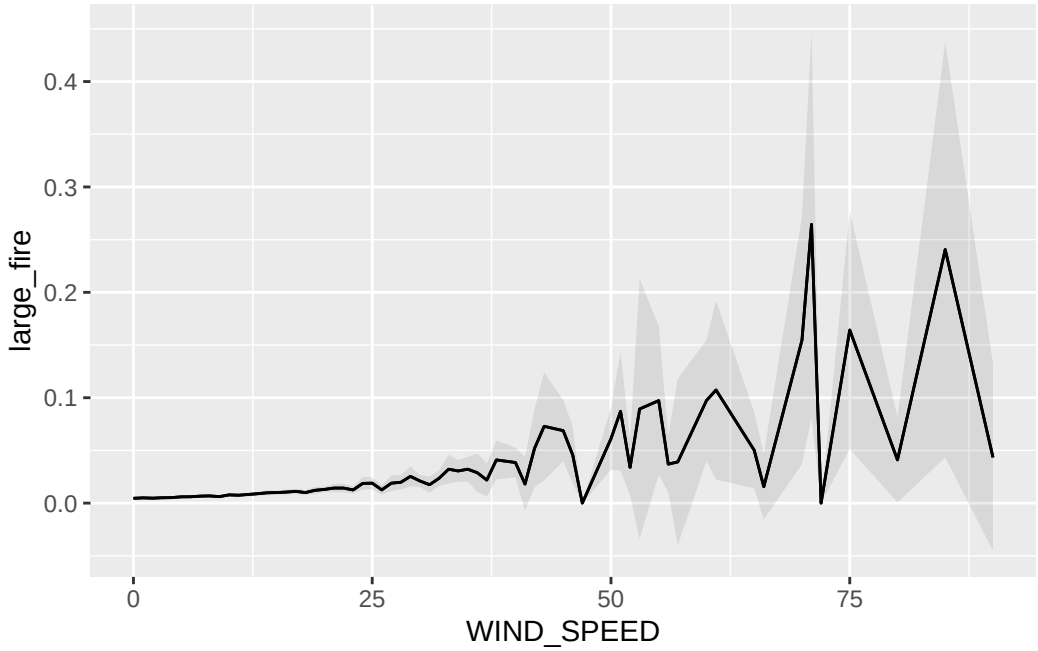
Interpretation Boost using `marginaleffects`

We used the `marginaleffects` package to interpret model predictions (Callaway and Arel-Bundock 2024).

Wind Speed

As wind speed increases, the model estimates a higher probability of a fire becoming large (>10 hectares). However, the variability in the predicted probabilities also increases at higher wind speeds, as indicated by the wider confidence intervals. This suggests that while there is a general upward trend, the model's certainty about the exact magnitude of the effect decreases in this range—likely due to fewer observations or greater variability in fire outcomes at high wind speeds.

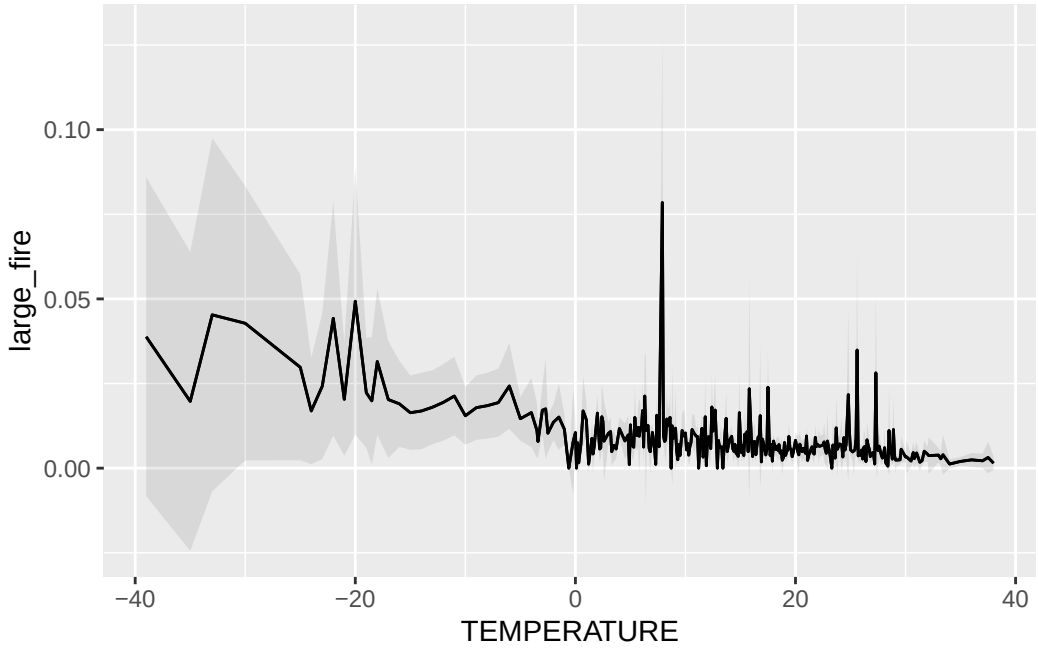
```
library(marginaleffects)
## continuous variable
plot_predictions(
  model,
  by = "WIND_SPEED"
)
```



Temperature

As temperature increases, the model predicts a relatively stable probability of a fire becoming large. The trend line flattens and the confidence intervals narrow, indicating that the model is more confident and consistent in its estimates across higher temperature ranges. This suggests that the relationship between temperature and fire size is more stable and predictable at higher temperatures, possibly due to a larger number of observations or less variability in outcomes.

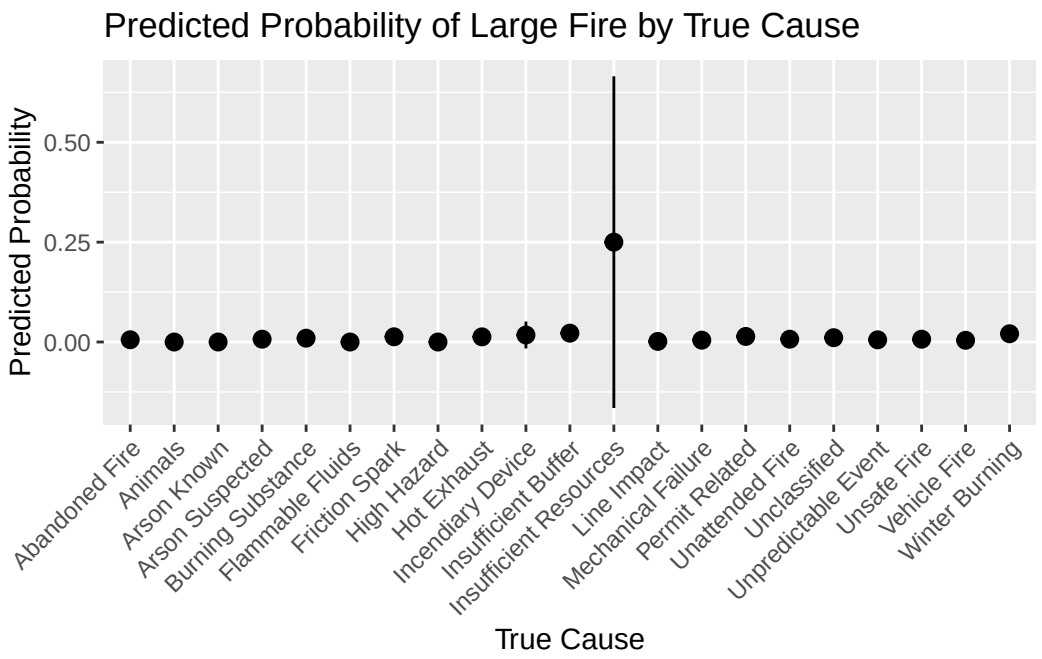
```
## continuous variable "TEMPERATURE"

plot_predictions(
  model,
  by = "TEMPERATURE"
)
```



True Cause

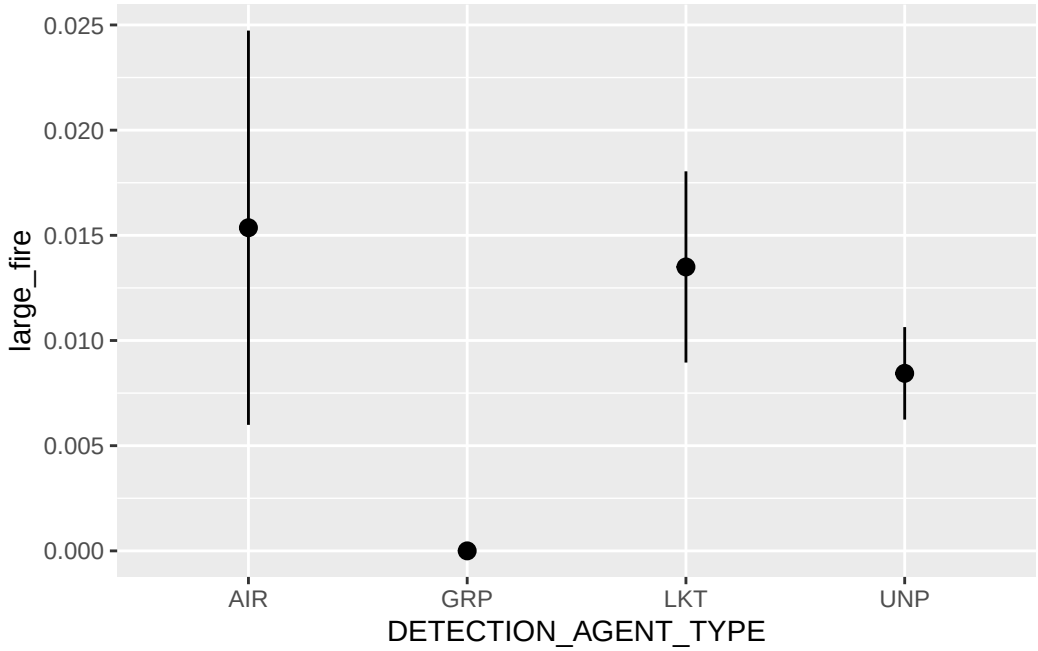
The predicted probability of a large fire is near zero for most `TRUE_CAUSE` categories, indicating that these causes (e.g., natural ignition, campfires, equipment use) are generally not associated with large-scale fires. However, the category **"Insufficient Resources"** stands out with a significantly higher predicted probability and a wide confidence interval. This suggests that fires classified under this cause are much more likely to become large, though the wide interval reflects substantial uncertainty — likely due to a small number of observations in that category.

```
## categorical variable "TRUE_CAUSE"
plot_predictions(model, by = "TRUE_CAUSE") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 45, hjust = 1)) +
  ggplot2::labs(
    title = "Predicted Probability of Large Fire by True Cause",
    x = "True Cause",
    y = "Predicted Probability"
  )
```



Detection Agent Type

```
## categorical variable "DETECTION_AGENT_TYPE"
plot_predictions(
  model,
  by = "DETECTION_AGENT_TYPE"
)
```



Attribution

- **Original Data Source**: Government of Canada – National Fire Database (NFDB)
- **License**: Open Government License – Canada
- **Citation**: Canadian Forest Service. National Fire Database (NFDB). Natural Resources Canada.
- **References**:

- `Logistic regression using `glm()` in R`
- `EDA best practices from Wickham & Grolemund (2017)`
- `Average Marginal Effects and Model Interpretation using `marginaleffects` (Vincent Arel-`

- **Acknowledgements**: Thanks to the Diverse Data Hub team for cleaning and contextualizing the data.

References

Callaway, Brantly, and Vincent Arel-Bundock. 2024. "Marginal Effects: Intuitive, Consistent, and Powerful Marginal Effects in r." https://marginaleffects.com.

GO FAIR Initiative. 2024. "The FAIR Data Principles." *GO FAIR*. https://www.go-fair.org/fair-principles/.

Wickham, Hadley, and Jenny Bryan. 2023. *R Packages*. O'Reilly Media. https://r-pkgs.org.