CDS 101 – Final Project Report

Samir Dawar, Andrew Lee, Sumera Muhamm

December 08, 2025

Contents

1. Problem Definition

The Chernobyl nuclear disaster released radioactive materials into the atmosphere, and these contaminants were carried across Europe by wind and weather systems. In this project, we study how the concentrations of three isotopes (Cs-137, Cs-134, and I-131) vary with distance from the Chernobyl Nuclear Power Plant. Our main goal is to see whether radiation levels decrease as distance increases and whether each isotope behaves differently due to differences in half-life. Cs-137 has a long half-life (about 30 years), Cs-134 decays faster, and I-131 decays very quickly (around 8 days). By comparing them, we can understand which isotopes show clear geographic decay and which are more influenced by timing and environmental conditions.

2. Data Acquisition & Description

We used the “Chernobyl Chemical Radiation / CSV / Country Data” dataset from Kaggle. After downloading the file, we saved it in our project folder under `data/Chernobyl_Chemical_Radiation.csv`.

The dataset contains: - Country code (`PAYS`) - Monitoring station name (`Location`) - Longitude and latitude - Date of measurement - Concentrations of I-131, Cs-134, and Cs-137 (Bq/m3)

Each row is one observation from a monitoring location on a specific day shortly after the accident. The dataset does not include any personal data, so there are no privacy concerns.

3. Data Cleaning & Preprocessing

We selected the variables needed for our analysis and renamed the isotope columns to shorter and easier names. Rows with missing isotope values were removed. To study how radiation changes with distance, we created a new variable called `distance` using a simple distance formula based on longitude and latitude differences.

```r
chernobyl <- read.csv("data/Chernobyl_ Chemical_Radiation.csv")

chernobyl_clean <- chernobyl %>%
  select(
    PAYS,
    Location,
    Longitude,
    Latitude,
    I_131 = I_131_.Bq.m3.,
    Cs_134 = Cs_134_.Bq.m3.,
    Cs_137 = Cs_137_.Bq.m3.
  ) %>%
  mutate(
    I_131 = as.numeric(I_131),
    Cs_134 = as.numeric(Cs_134),
    Cs_137 = as.numeric(Cs_137)
  ) %>%
  filter(!is.na(Cs_137))

#distrance measurement
chernobyl_clean <- chernobyl_clean %>%
  mutate(
    distance = sqrt((Longitude - 30.099)^2 + (Latitude - 51.389)^2)
  )
head(chernobyl_clean)
```

```
##   PAYS Location Longitude Latitude   I_131  Cs_134  Cs_137 distance
## 1   SE    RISOE     12.07     55.7 1.00000 0.00000 0.24000 18.53725
## 2   SE    RISOE     12.07     55.7 0.00460 0.00054 0.00098 18.53725
## 3   SE    RISOE     12.07     55.7 0.01470 0.00430 0.00740 18.53725
## 4   SE    RISOE     12.07     55.7 0.00061 0.00000 0.00009 18.53725
## 5   SE    RISOE     12.07     55.7 0.00075 0.00010 0.00028 18.53725
## 6   SE    RISOE     12.07     55.7 0.00053 0.00000 0.00020 18.53725
```

4. Exploratory Data Analysis (EDA)

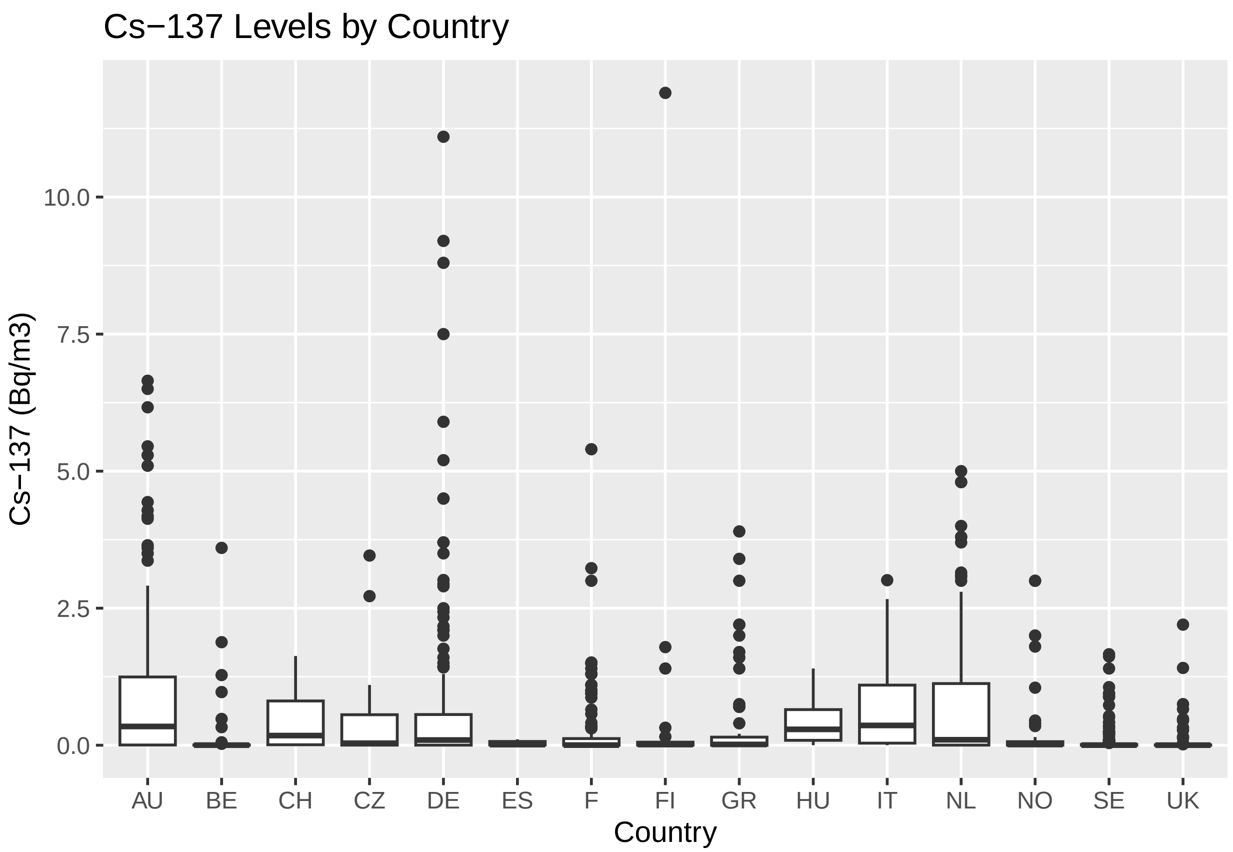
The goal of this section is to explore the dataset and understand how radiation levels differ across countries and how they change with distance from the Chernobyl plant. We look at summary statistics, boxplots, scatterplots, and a geographic-style plot to visually assess the relationships that will later be modeled. We begin by examining simple descriptive statistics for the three isotopes.

```
summary(chernobyl_clean[, c("I_131", "Cs_134", "Cs_137")])
```

```
##      I_131               Cs_134            Cs_137
##  Min.   : 0.00000   Min.   :0.00000   Min.   : 0.0000
##  1st Qu.: 0.00350   1st Qu.:0.00000   1st Qu.: 0.0016
##  Median : 0.05925   Median :0.00197   Median : 0.0200
##  Mean   : 1.43874   Mean   :0.21225   Mean   : 0.4801
##  3rd Qu.: 1.03016   3rd Qu.:0.14100   3rd Qu.: 0.4793
##  Max.   :27.59828   Max.   :7.20000   Max.   :11.9000
##  NA's   :6          NA's   :85
```

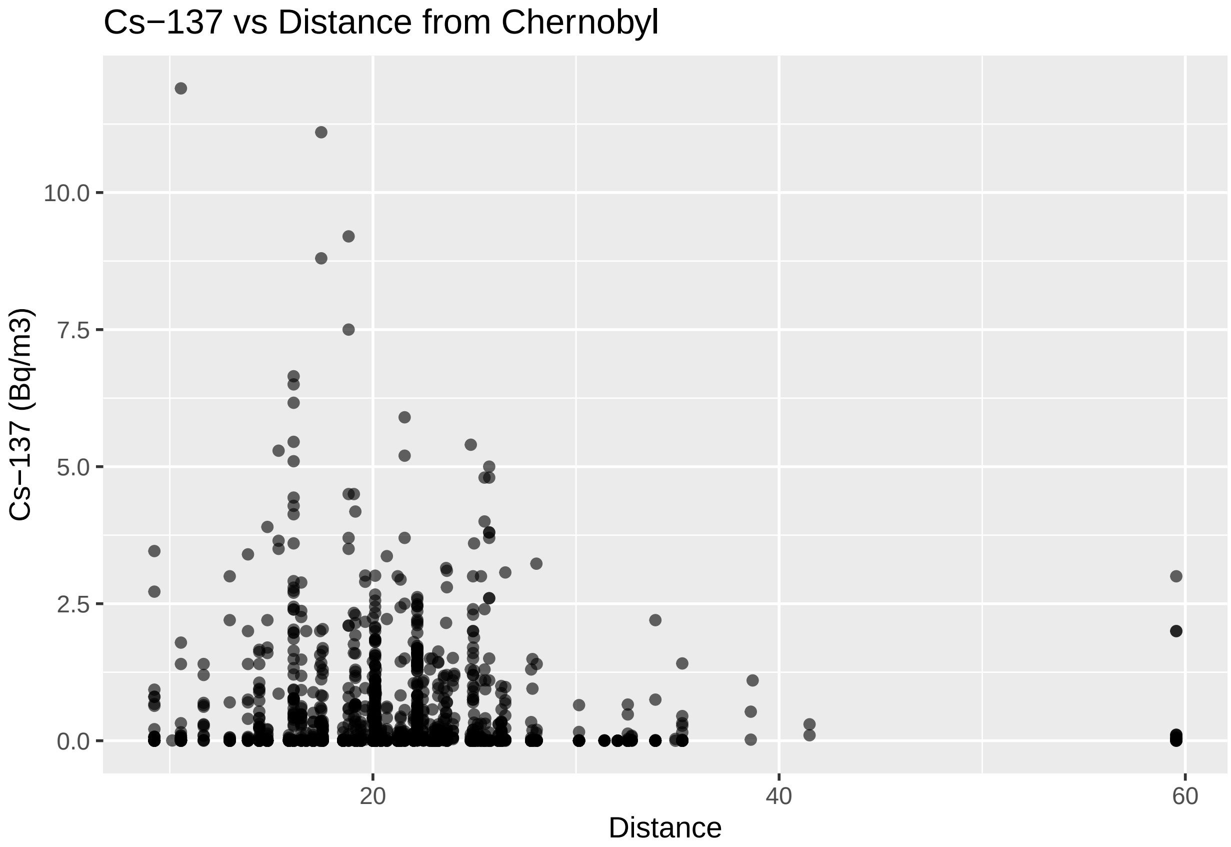
The summary statistics show that all three isotopes have very skewed distributions. Most values are close to zero, but there are a few large measurements, especially for Cs-137 and I-131. This is expected because only certain monitoring stations recorded heavy fallout. Cs-134 has more missing values, which reflects gaps in collection across countries. These summaries help us understand the starting point before plotting.

```
ggplot(chernobyl_clean, aes(x = PAYS, y = Cs_137)) +
geom_boxplot() +
labs(
title = "Cs-137 Levels by Country",
x = "Country",
y = "Cs-137 (Bq/m3)"
)
```



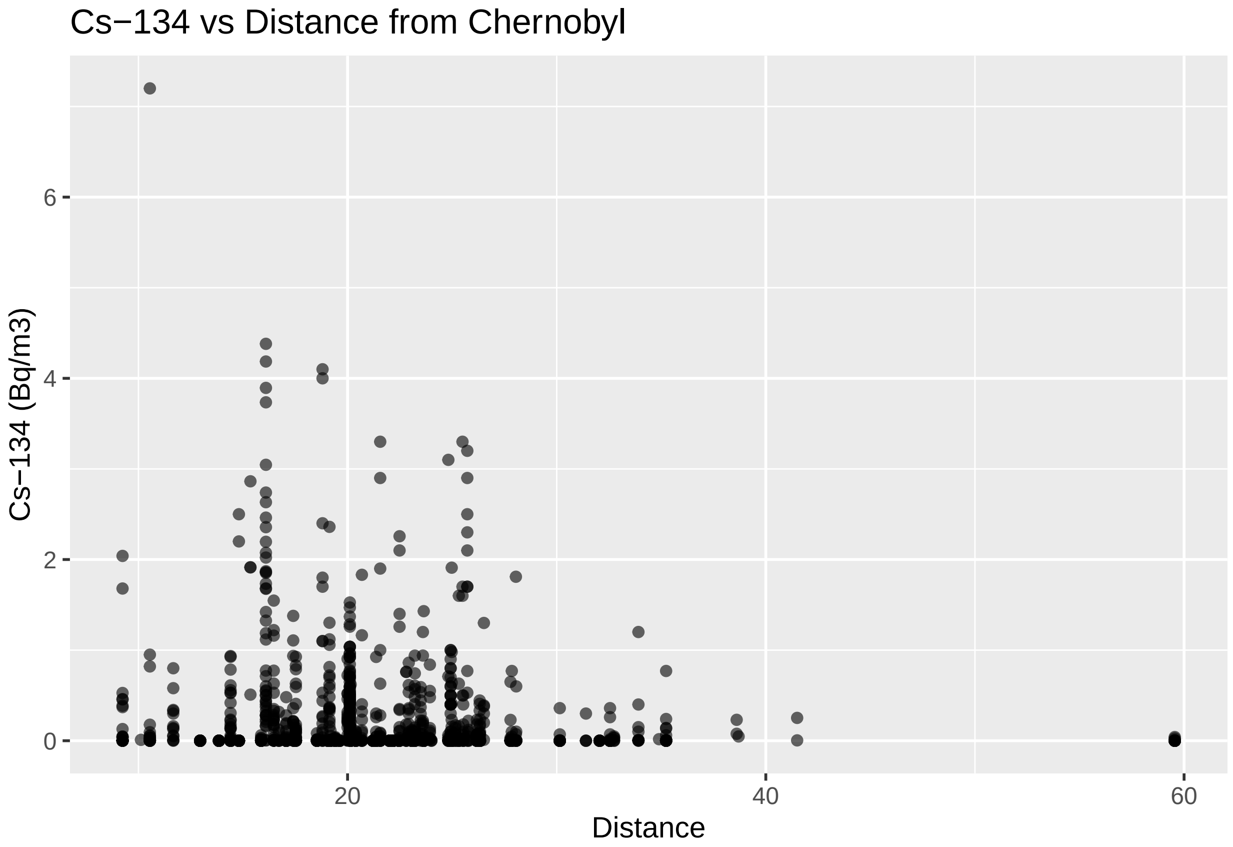
This boxplot shows that Cs-137 levels vary noticeably across different countries. Some countries have higher medians and more extreme outliers, while others have consistently low measurements. Because Cs-137 is a long-lasting isotope, these differences likely reflect how fallout patterns moved across Europe after the explosion. This supports the idea that geographic location influenced radiation exposure levels.

```
ggplot(chernobyl_clean, aes(distance, Cs_137)) +
geom_point(alpha = 0.6) +
labs(
title = "Cs-137 vs Distance from Chernobyl",
x = "Distance",
y = "Cs-137 (Bq/m3)"
)
```



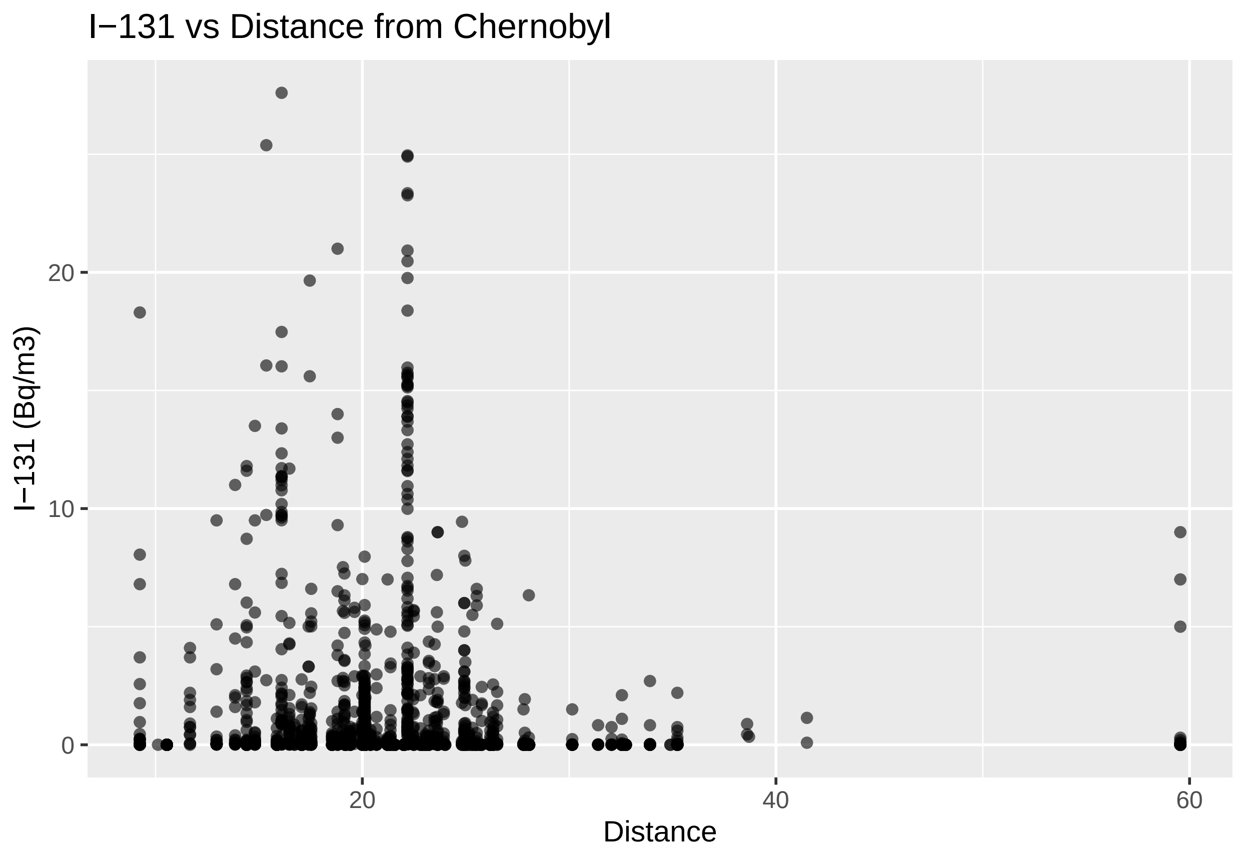
Cs-137 clearly decreases as distance from Chernobyl increases. Locations closer to the plant have higher concentrations and more extreme values, while farther locations cluster near zero. This matches scientific expectations because Cs-137 persists in the environment and spreads outward with decreasing intensity. This visualization directly supports our main research question.

```r
ggplot(chernobyl_clean, aes(distance, Cs_134)) +
geom_point(alpha = 0.6) +
labs(
title = "Cs-134 vs Distance from Chernobyl",
x = "Distance",
y = "Cs-134 (Bq/m3)"
)
```



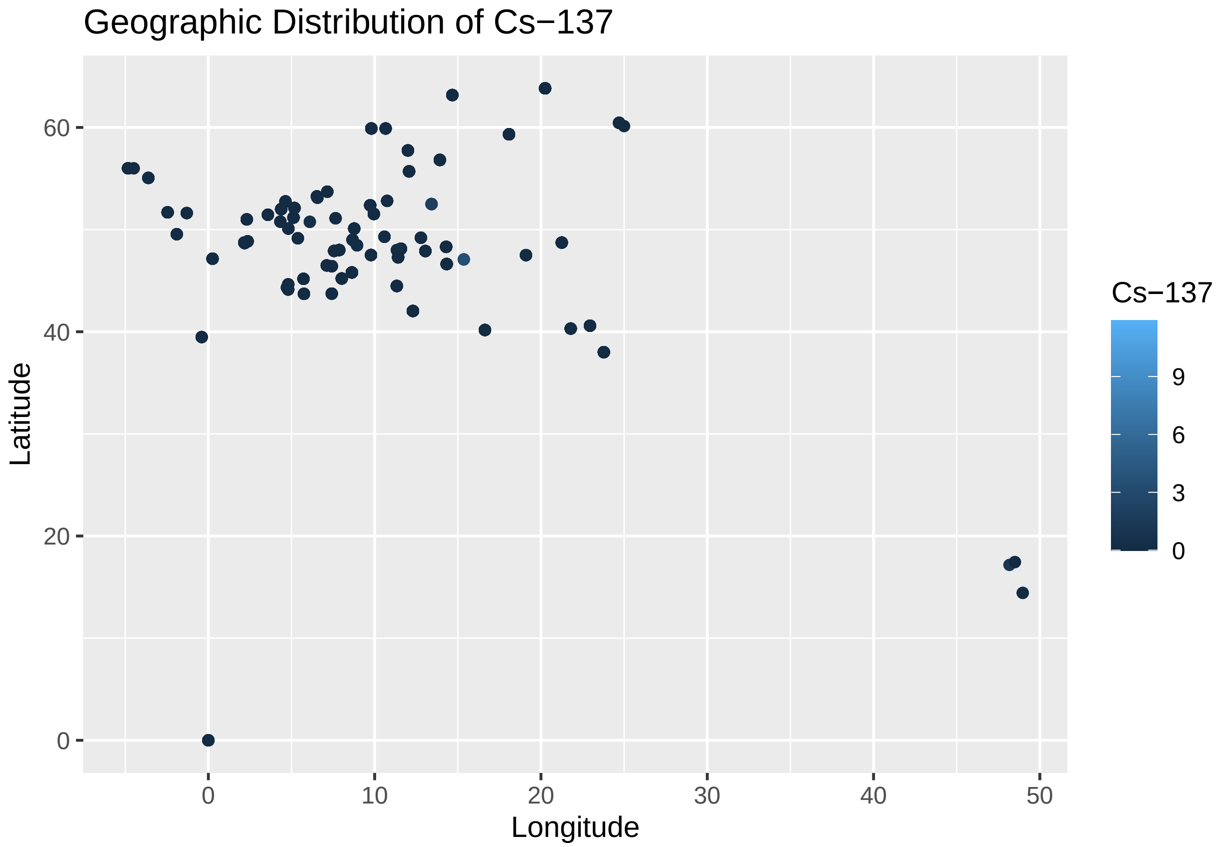
Cs-134 also shows a downward pattern with distance, but the trend is weaker than Cs-137. Many values are close to zero, and there is more variation overall. This is expected because Cs-134 decays faster than Cs-137, so its measurements depend more on timing and local environmental conditions. Still, the overall pattern suggests that distance plays some role in Cs-134 concentration.

```
ggplot(chernobyl_clean, aes(distance, I_131)) +
geom_point(alpha = 0.6) +
labs(
title = "I-131 vs Distance from Chernobyl",
x = "Distance",
y = "I-131 (Bq/m3)"
)
```



I-131 does not show a clear relationship with distance. Measurements are scattered at many distances, including some high values that appear far from the plant. This behavior makes sense because I-131 has a very short half-life (about eight days), so its concentration depends more on the date of measurement and weather patterns than on geographic distance. Distance alone cannot explain I-131 levels.

```r
ggplot(chernobyl_clean, aes(Longitude, Latitude, color = Cs_137)) +
geom_point() +
labs(
title = "Geographic Distribution of Cs-137",
x = "Longitude",
y = "Latitude",
color = "Cs-137"
)
```



This map-style plot shows where the highest Cs-137 values occurred across Europe. Higher concentrations appear in areas closer to the Chernobyl plant and in regions where prevailing winds carried the fallout. This plot reinforces that contamination was not uniform across Europe and supports the idea that both distance and location influenced radiation levels.

5. Visualization Quality and Storytelling

Our goal is to measure how radiation levels change with distance from the Chernobyl Nuclear Power Plant. Because our response variables (Cs-137, Cs-134, and I-131) are numeric measurements, this is a regression task. Based on our Exploratory Data Analysis, Cs-137 shows the clearest decreasing trend with distance, so we use it as the main variable for modeling. We use a simple linear regression model with Cs-137 as the response variable and distance as the predictor. A linear model is appropriate here because: It is designed for continuous outcomes. It provides an interpretable slope that shows whether Cs-137 decreases with distance. It matches the roughly linear downward pattern we observed in the scatterplot. This model allows us to quantify the direction and strength of the relationship between distance and radiation levels, directly addressing our research question about how contamination decreased across Europe.

6. Modeling Approach

Explain how you framed the problem and which models you chose:

- Type of task (regression, classification, etc.)
- Baseline model or heuristic, if used.

- Main model(s) chosen and why they are appropriate.

7. Model Implementation & Evaluation

To evaluate how well distance explains radiation levels, we fit a simple linear regression model using Cs-137 as the response variable and distance as the predictor. Cs-137 was chosen because it showed the clearest decreasing pattern with distance in our EDA.

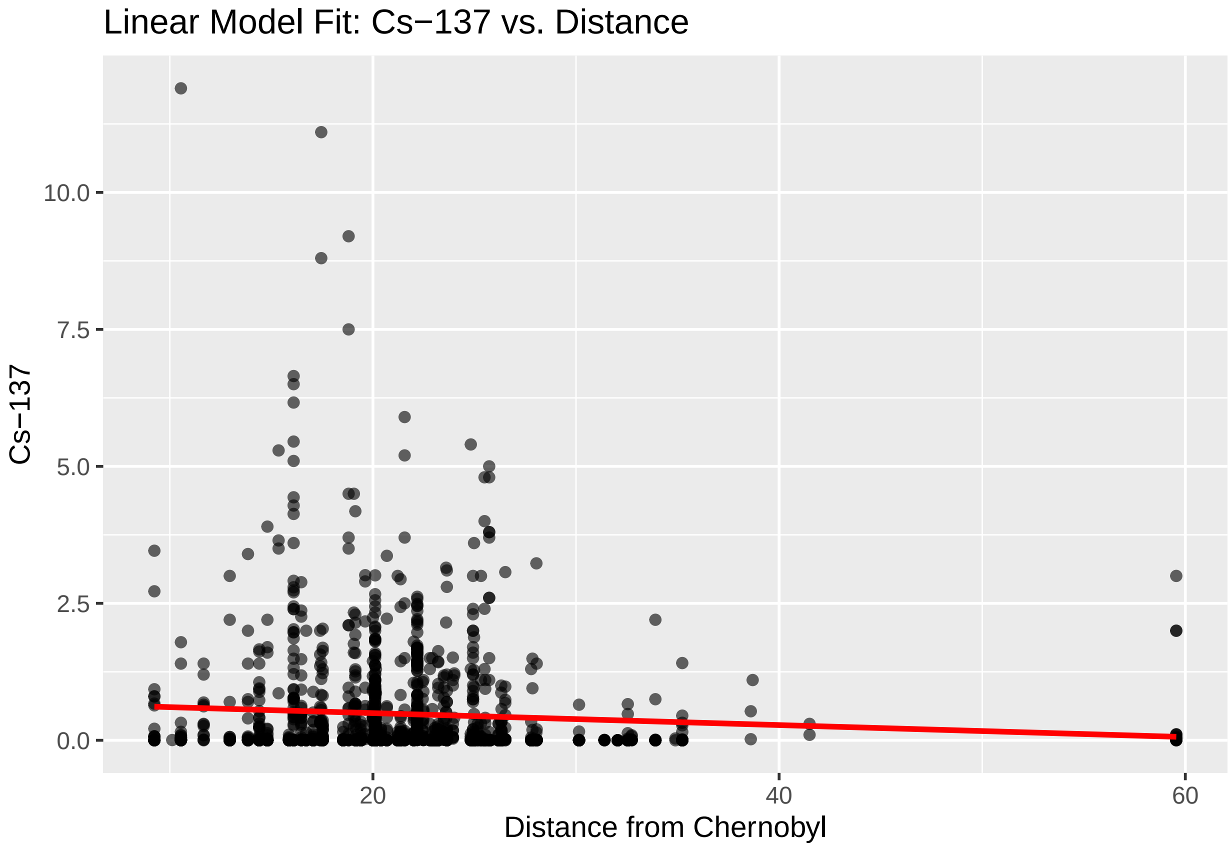
A linear regression is appropriate because both variables are numeric and our goal is to measure how Cs-137 changes for every unit increase in distance from the plant.

```
model_cs137 <- lm(Cs_137 ~ distance, data = chernobyl_clean)
summary(model_cs137)
```

```
##
## Call:
## lm(formula = Cs_137 ~ distance, data = chernobyl_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6123 -0.4921 -0.4044  0.0009 11.3007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.714363   0.089008   8.026 2.01e-15 ***
## distance    -0.010914   0.003957  -2.758  0.00588 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.034 on 1504 degrees of freedom
## Multiple R-squared:  0.005033,   Adjusted R-squared:  0.004371
## F-statistic: 7.607 on 1 and 1504 DF,  p-value: 0.005883
```

```
ggplot(chernobyl_clean, aes(distance, Cs_137)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Linear Model Fit: Cs-137 vs. Distance",
    x = "Distance from Chernobyl",
    y = "Cs-137"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



8. Conclusions & Recommendations

Our main research question was whether radiation level decreased as distance from the Chernobyl Nuclear Power Plant increased, and whether different isotopes behaved differently because of their half-lives.

Key Findings:

- Cs-137 shows the clearest distance-based pattern. In both the scatterplot and the linear model, Cs-137 decreases as distance increases. The trend is small but statistically significant. This fits what we expect from a long-lasting isotope that traveled across Europe with the wind but still shows a general decline farther from the disaster site.
- Cs-134 also decreases with distance but more weakly. Many Cs-134 values were close to zero, and the relationship was noisier. Because Cs-134 decays faster than Cs-137, the amount measured depended more on timing and local conditions, which adds variability.
- I-131 does not follow a clear distance pattern. The scatterplot showed high scatter at all distances. Since I-131 has a short half-life of about eight days, its concentration depends heavily on when each measurement was taken rather than how far the location was from Chernobyl. Distance alone cannot explain I-131 levels.
- The linear regression model supports the EDA patterns. For Cs-137, the model found a negative slope, meaning levels decrease as distance increases. However, the R^2 value was small, showing that distance

explains only a small part of the variation. Many other factors also affected fallout levels, such as wind direction, storms, time of sampling, and geography.

Limitations

- Some countries submitted many measurements while others had very few. This affects how well the model can generalize.
- Without the exact date of each sample, we cannot accurately model fast-decaying isotopes like I-131.
- We used an approximate distance formula based on longitude and latitude. A more precise method or real geographic distance would improve accuracy.
- Rainfall, wind direction, and terrain played huge roles in fallout patterns, but the dataset does not include these variables.

Recommendations & Future Work

- If measurement dates were available, we could model decay curves and understand how quickly concentrations dropped over time.
- Including latitude, longitude, and possibly direction from the reactor could improve prediction models.
- Radiation fallout rarely decreases in a perfectly linear way. A curved or spatially-aware model could capture patterns better.
- A combined model could help show how half-life affects the relationship between distance and contamination levels.

Distance from Chernobyl does play a role in radiation levels, especially for long-lasting isotopes like Cs-137. However, distance alone cannot fully explain the variation we see across Europe. Fallout patterns were shaped by many factors, and radiation behavior differs strongly depending on the isotope. Our results reinforce both the scientific understanding of the disaster and the importance of considering physical and environmental processes when modeling real-world data.

9. Code Quality & Reproducibility

1. Download the dataset below and place it in the data folder https://www.kaggle.com/datasets/brsdincer/chernobyl-chemical-radiation-csv-country-data
2. Open the R Markdown file in Rstudio titled final_report_template.Rmd
3. install the required R packages if they are not already installed by running the following commands in the console.

```
install.packages("dplyr")
install.packages("ggplot2")
```

4. Run each code block in the document and finally Knit.

```
## R version 4.4.1 (2024-06-14)
## Platform: aarch64-apple-darwin20
## Running under: macOS 26.0
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_4.0.1 dplyr_1.1.4
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.5        nlme_3.1-164       cli_3.6.5          knitr_1.50
##  [5] rlang_1.1.6        xfun_0.53          generics_0.1.4     S7_0.2.1
##  [9] labeling_0.4.3     glue_1.8.0         htmltools_0.5.8.1  scales_1.4.0
## [13] rmarkdown_2.29     grid_4.4.1         evaluate_1.0.4     tibble_3.3.0
## [17] fastmap_1.2.0      yaml_2.3.10        lifecycle_1.0.4    compiler_4.4.1
## [21] RColorBrewer_1.1-3 pkgconfig_2.0.3    mgcv_1.9-1         rstudioapi_0.17.1
## [25] lattice_0.22-6     farver_2.1.2       digest_0.6.37      R6_2.6.1
## [29] tidyselect_1.2.1   splines_4.4.1      pillar_1.11.0      magrittr_2.0.3
## [33] Matrix_1.7-0       withr_3.0.2        tools_4.4.1        gtable_0.3.6
```

10. References

- Kaggle dataset used for this project -> https://www.kaggle.com/datasets/brsdincer/chernobyl-chemical-radiation-csv-country-data
- https://www.epa.gov/radiation/radiation-terms-and-units
- https://www.r-project.org/
- https://www.sciencedirect.com/topics/medicine-and-dentistry/cesium-134#:~:text=Cesium%2D134%20is%20a%20rad
- https://www.cdc.gov/radiation-emergencies/hcp/isotopes/cesium-137.html
- https://www.oecd-nea.org/jcms/pl_28351/chernobyl-chapter-vi-agricultural-and-environmental-impacts