HW 6 (Reproducible Reporting)

Laavanya Joshi Malik

2023-10-10

Introduction: For my reproducible reporting project, I'm using the Lead-IQ data set which has been derived as a result of the original study published, Neuro-Psychological dysfunction in children with chronic low-level lead absorption in 1975 by Landrigan PJ, Baloh RW, Barthelt WF, Whitworth RH, Staehlinh NW, and Rosenblum BF.

I began with installing the packages I need and then downloaded the libraries.

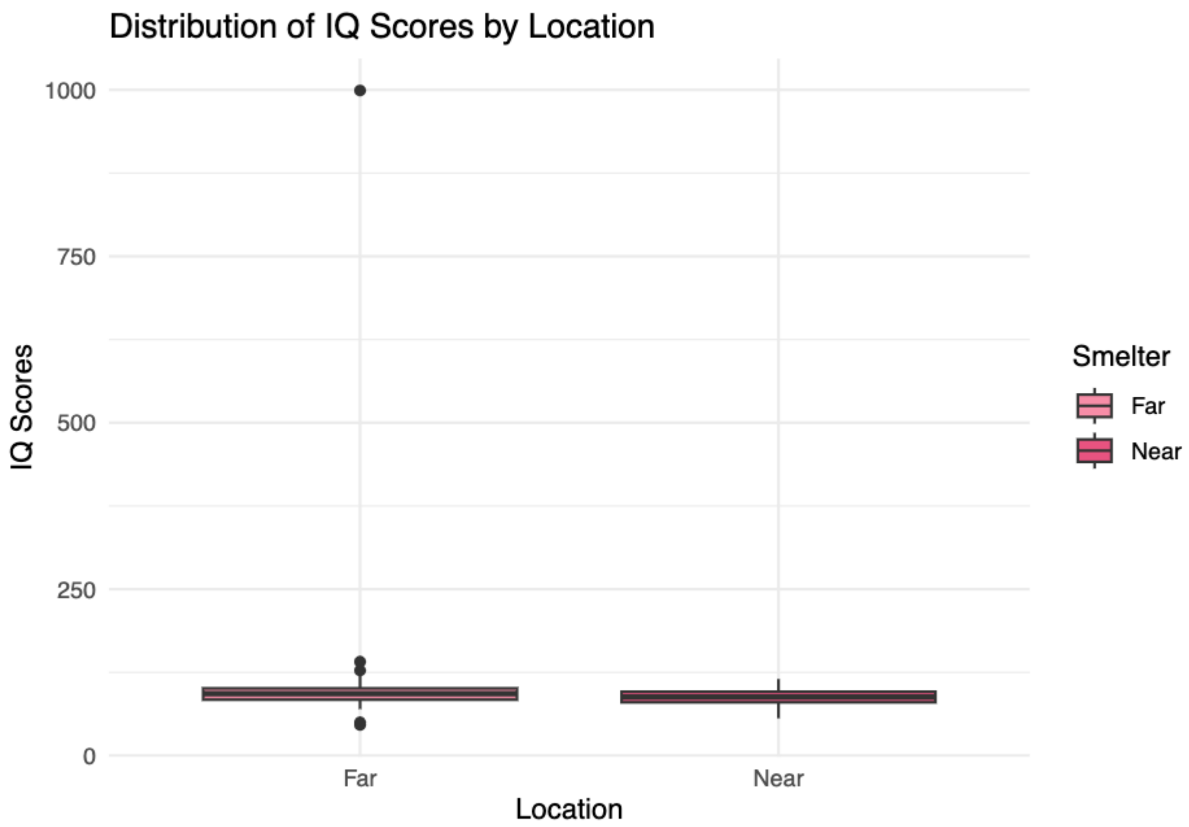
Uploading data set and creating data frame: After utilizing tidyverse library to upload the dataset, I also created my data frame, lead. This organization will help me with my exploratory analysis.

```
file_path <- "/Users/Laavanya/Malik_Project_01/DataRaw/lead-iq-01.csv"
lead <- read_csv(file_path)

#head(lead)
#tail(lead)
```

For part A: Graph showcases mean IQ levels by location status:

```
ggplot(lead, aes(x = Smelter, y = IQ, fill = Smelter)) +
  geom_boxplot() +
  scale_fill_manual(values=c("Far" = "#F78DA7", "Near" = "#E75480")) +
  labs(title = "Distribution of IQ Scores by Location",
       x = "Location",
       y = "IQ Scores") +
  theme_minimal()
```



Interpretation (part C):

Box Plot: The boxplot displays the distribution of IQ score for children that live "Far" and "Near" the large, lead-emitting ore smelter. There are some outliers for the IQ scores for the "Far" which are WAY above the maximum possible IQ score. This indicates a flaw in the data. Nonetheless, the boxplot helps a lot with helping us see the defect/inaccuracy.

For part B: Table with subset of values from Lead-IQ data set!

```r
even_iq_data <- lead %>%
  filter(IQ %% 2 == 0, IQ > 90)

even_iq_table <- even_iq_data %>%
  kable("simple") %>%
  kable_styling("striped", full_width = F)

# Print the table
even_iq_table
```

Smelter	IQ
Far	96
Far	94
Far	128
Far	118
Far	96
Far	96
Far	120
Far	100
Far	94
Far	94
Far	104
Far	92
Far	100
Far	98
Far	104
Far	96
Far	108
Far	102
Far	92
Far	92
Near	96
Near	96
Near	106
Near	98
Near	104
Near	96
Near	94
Near	104
Near	112
Near	92
Near	114
Near	96

Interpretation (part C)

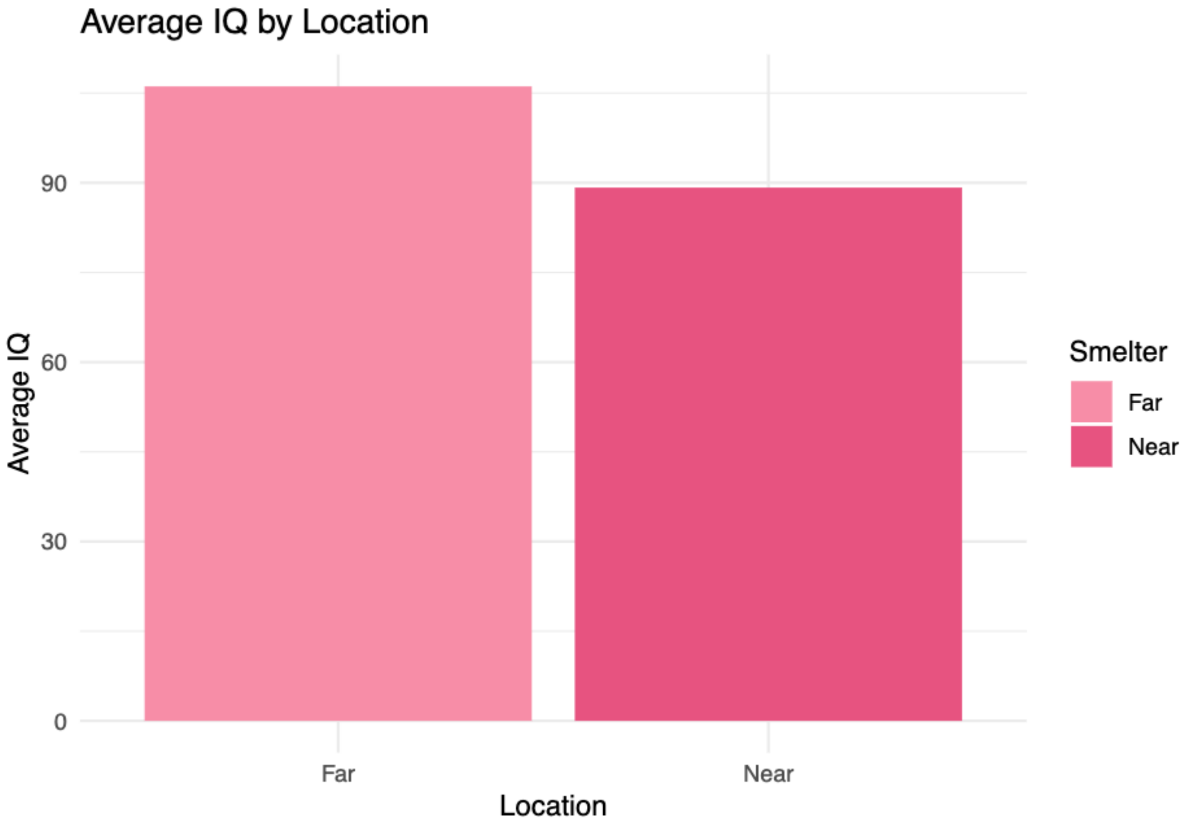
Table: Since the data set is 124 observations in total, I decided to create a table with just the IQ scores that are even and greater than 90. It's a simple table style (although, I did install the kableExtra package).

For part D:

Calculating the means of the IQ scores. Further, I also decided to create a box plot with the average IQ scores per location type, which is binary.

```
mean_iq_near <- mean(lead$IQ[lead$Smelter == "Near"], na.rm = TRUE)
mean_iq_far <- mean(lead$IQ[lead$Smelter == "Far"], na.rm = TRUE)

lead %>%
  group_by(Smelter) %>%
  summarise(MeanIQ = mean(IQ)) %>%
  ggplot(aes(x = Smelter, y = MeanIQ, fill = Smelter)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=c("Far" = "#F78DA7", "Near" = "#E75480")) +
  labs(title = "Average IQ by Location",
       x = "Location",
       y = "Average IQ") +
  theme_minimal()
```



The avg IQ score for near the smelter is 89.2 and far from the smelter is 106.1.

For part 3

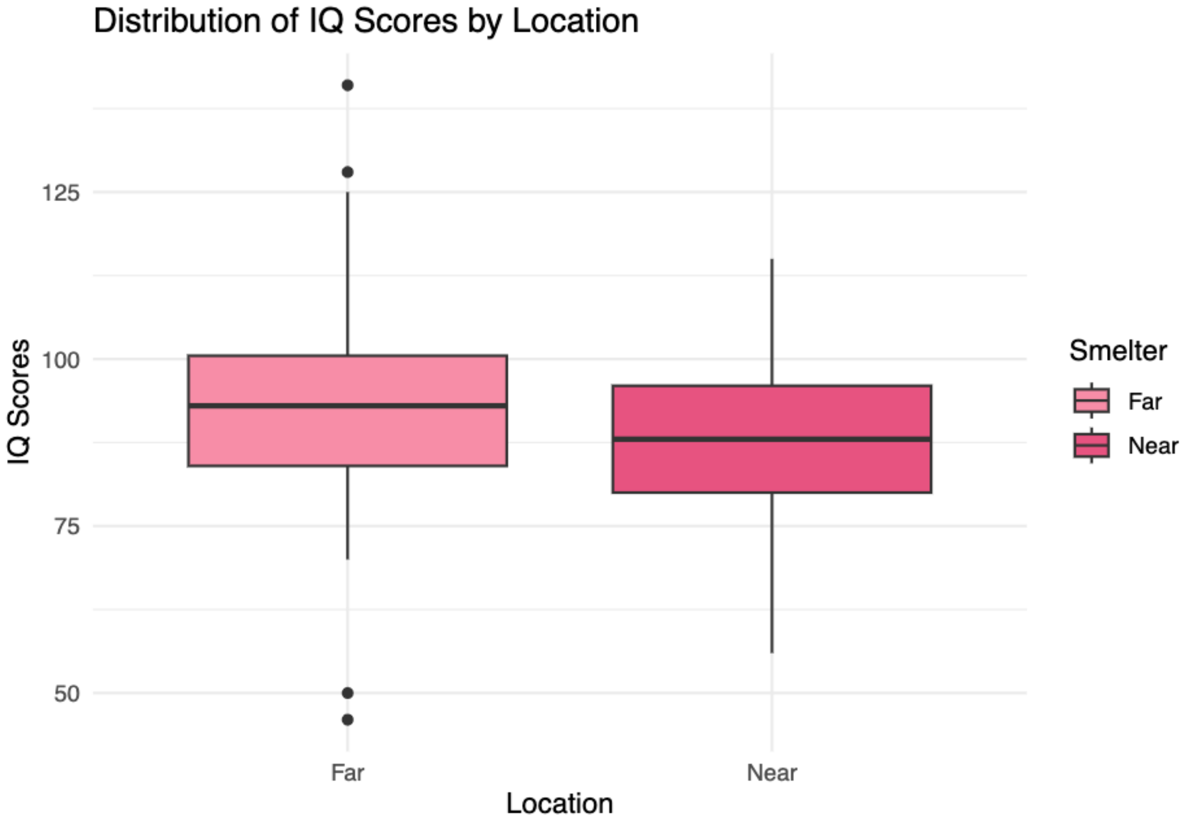
I did notice something weird. I definitely caught this as soon as I noticed the box plot in part A where I was simply displaying my data set. The value 999 is extremely odd and did give me the hint that it was supposed to be 99.

```r
lead_mod <- lead
lead_mod$IQ[lead_mod$IQ == 999] <- 99

write.csv(lead_mod, file = "corrected_dataset.csv", row.names = FALSE)
```

For part A

```r
ggplot(lead_mod, aes(x = Smelter, y = IQ, fill = Smelter)) +
  geom_boxplot() +
  scale_fill_manual(values=c("Far" = "#F78DA7", "Near" = "#E75480")) +
  labs(title = "Distribution of IQ Scores by Location",
       x = "Location",
       y = "IQ Scores") +
  theme_minimal()
```



Interpretation (part C):

Box Plot: The boxplot displays the distribution of IQ score for children that live “Far” and “Near” the large, lead-emitting ore smelter. As we can see, the median IQ for “Far” is greater than the IQ for “Near.” There are some outliers for the IQ scores for the “Far” showcasing that these individuals’ IQ scores may be impacted by other circumstances, which seems right since they live “far” from the lead-emitting ore smelter.

Nonetheless, there isnt great variation between the IQ scores, which can mean there are socioeconomic and educational factors that influence our data set.

For part B

```r
even_iq_data <- lead_mod %>%
  filter(IQ %% 2 == 0, IQ > 90)

even_iq_table <- even_iq_data %>%
  kable("simple") %>%
  kable_styling("striped", full_width = F)

# Print the table
even_iq_table
```

Smelter	IQ
Far	96
Far	94
Far	128
Far	118
Far	96
Far	96
Far	120
Far	100
Far	94
Far	94
Far	104
Far	92
Far	100
Far	98
Far	104
Far	96
Far	108
Far	102
Far	92
Far	92
Near	96
Near	96
Near	106
Near	98
Near	104
Near	96
Near	94
Near	104
Near	112
Near	92
Near	114
Near	96

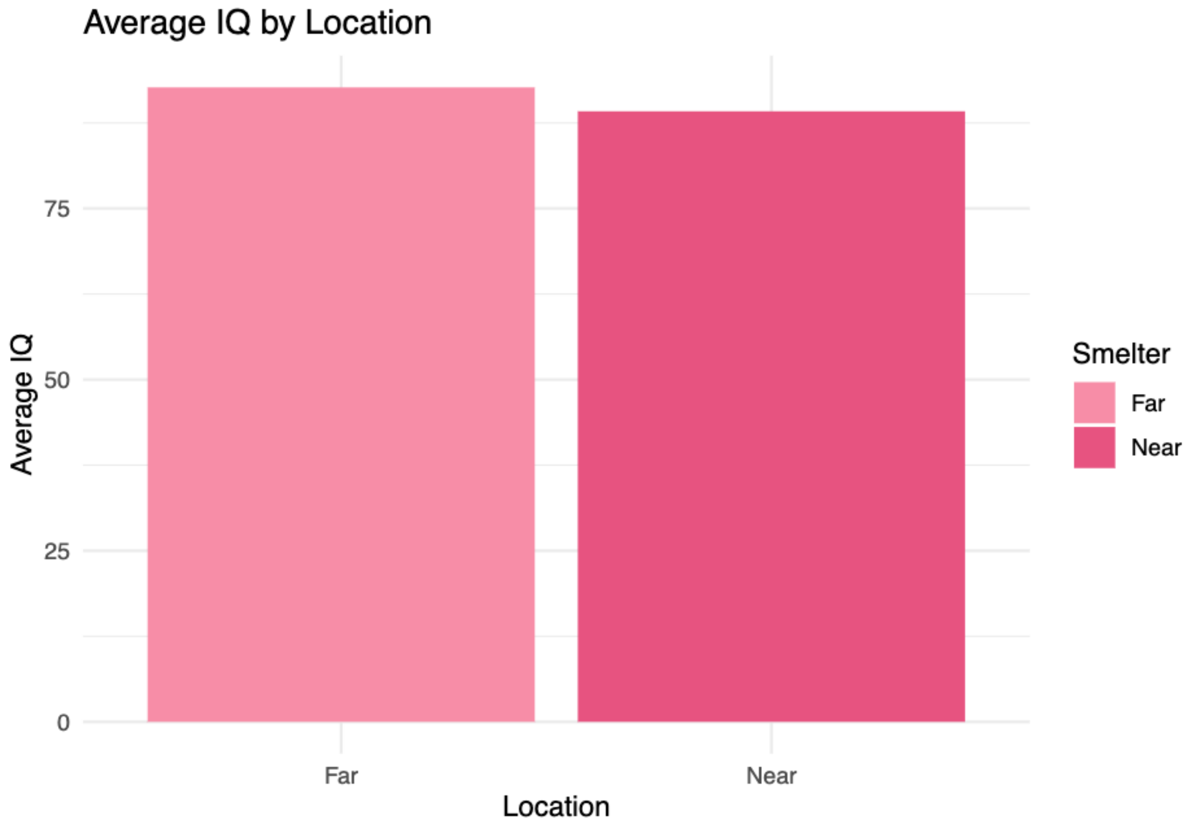
Interpretation (part C)

Table: We don't notice a big change to this since my subset of data was even AND greater than 90. 99 is an odd number :)

For part D

```
mean_iq_near_mod <- mean(lead_mod$IQ[lead_mod$Smelter == "Near"], na.rm = TRUE)
mean_iq_far_mod <- mean(lead_mod$IQ[lead_mod$Smelter == "Far"], na.rm = TRUE)

lead_mod %>%
  group_by(Smelter) %>%
  summarise(MeanIQ = mean(IQ)) %>%
  ggplot(aes(x = Smelter, y = MeanIQ, fill = Smelter)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=c("Far" = "#F78DA7", "Near" = "#E75480")) +
  labs(title = "Average IQ by Location",
       x = "Location",
       y = "Average IQ") +
  theme_minimal()
```



The avg IQ score for near the smelter is 89.2 and far from the smelter is 92.7. The change we notice is the value of mean IQ score for "Far" children from the smelter.

Conclusion

This was a pretty fascinating data set to work with. It was very interesting to measure and explore the influence of harmful materials such as lead have on the human body.