Inspecting and understanding our data (and its metadata)

1. Is the data set a sample or a population? Why?

The data set is a population, because it includes all Health Department inspections for food service establishments in King County, which means it includes all members from the specified group of inspections.

2. What type of data structure is your variable Inspection.Score?

It is a matrix of integers

3. Why can we calculate more measures of central tendency for Inspection.Score compared to Inspection.Result? Which measures of central tendency could be calculated for Inspection.Result?

Inspection.Score uses numeric values, which can be used to calculate all measures of central tendency. Inspection.Result on the other hand uses characters, which can only be used to calculate median and mode.

4. What is the unit of Inspection.Score?

The units are numeric values that are on a scale of food safety violations

5. What does a high value of Inspection.Score correspond to in reality? What about a low value?

A high value of Inspection.Score corresponds to more violations and critical violations, whereas a low value corresponds to less violations and critical violations. 0 is a perfect score

6. What is the period for which food inspection data exists? (Hint: You can look through the data and refer to the website for further data about the data, or metadata)

The dataset has data starting from 01/01/2006 and goes to 10/13/2022

Calculating measures of central tendency for Inspection.Score

7. What is the mean?

14.127

8. What is the median?

5

9. Are the mean and median similar, or is one value larger than the other? What does any similarity or difference about Inspection.Score reveal?

The mean and median are not similar, with the mean being larger than the median. This reveals that the middle of the Inspection.Score values sorted is 5, which is much lower than the mean of Inspection.Score values, making the distribution of the data skewed to the right.

10. What is the range? In your own words, what does the range tell us?

max = 178, min = -30, range = 208

This tells us the largest possible difference that can be between values in the Inspection.Score column is 208.

11. What is the variance? In your own words, what does the variance tell us?

Variance = 412.494

This tells us the average squared distance from the mean of the values in Inspection.Score is 412.949.

12. What is the standard deviation of Inspection.Score? In your own words, what does the standard deviation reveal?

Sd = 20.310

This tells us the variability from the mean of the values in Inspection.Score is 20.310

Subsetting data to a specific year

We will now compare food inspection data between three different years: 2009, 2019, and 2020. We will do this by subsetting data based on the column Inspection.Date. *Note that I already converted this variable to a date format in Excel, so if you were to download the csv file anew from the King County website, you would need to convert the date column to that format either in Excel or R before proceeding.*)

First, let's subset the data to 2006:

```
df2006 <- subset(mydata, Inspection.Date >= "2006-01-01" & Inspection.Date <=
"2006-12-31")
```

Then, let's subset the data to 2019. Change the necessary values in the command above to do so.

13. How many inspections were carried out in 2006 versus 2019? Based on the values, does it make sense to compare data from these two years?

Inspections 2006 = 9,910, Inspections 2019 = 20,898

These years are over 10 years apart and have a difference of over 10,000 inspections. Since the values have such a difference, it makes sense to compare the years to find more differences and similarities so that we can notice changes and trends over time and as the volume of inspections increase.

14. What is the mean of Inspection.Score in 2006 versus 2019? What does any difference you notice suggest?

mean 2006 = 12.922, mean 2019 = 12.196

These means are relatively close, with the mean in 2019 being slightly lower. This tells us that the average inspection score in 2019 is slightly better than the average inspection score in 2006.

15. What is the standard deviation of Inspection.Score in 2006 versus 2019? What does any difference you notice suggest?

sd 2006 = 18.826

sd 2019 = 17.870

The standard deviation from 2019 is about 1 less than the standard deviation from 2006. This tells us that there is a lower variability from the mean in 2019's Inspection.Score than 2006's

Finally, subset the data to 2020.

How many inspections were carried out? What is the cause for this number, and would it make sense to compare data from this year with the previous two years?

Inspections 2020 = 7451

This number is much lower than that of 2019, which would usually be an unlikely difference to see in the matter of a year. This is because many food establishments were closed for the majority of 2020 due to the COVID-19 pandemic. I think it would make sense to compare this to the other years, as it would be interesting to see how food establishments during the pandemic score against years before the pandemic.

Reflections

17. What was the most challenging aspect of this assignment for you, if any?

It wasn't too challenging, but it was harder for me to understand the data in terms of things like its data structure.

18. What other analysis would you be interested in running on this dataset?

I'd like to compare data from 2019 to data from 2021, to see the difference between number of inspections and scores pre pandemic and post lockdown pandemic to see how they differ. I am interested to see if food establishments tend to do better or worse, which could be affected by potential lack in staffing or potential increase in sanitary practices.

Before your section meets on Wednesday, October 19, please submit a single PDF document on Canvas that includes 1) your answers to questions 1-18 and 2) your commands in the RStudio Console (these can be copied and pasted into a word processing document underneath the answers to your questions.)

```
> setwd('/Users/My-An/Docs/geog317/assignment2')

> mydata <- read.csv('food_inspection.csv')

> mean(mydata$Inspection.Score, na.rm = TRUE)

[1] 14.12698

> median(mydata$Inspection.Score, na.rm = TRUE)

[1] 5

> range(mydata$Inspection.Score, na.rm = TRUE)

[1] -30 178

> 178-(-30)

[1] 208

> var(mydata$Inspection.Score, y=NULL, na.rm = TRUE)

[1] 412.4938

> sd(mydata$Inspection.Score, na.rm = TRUE)

[1] 20.30994

> df2006 <- subset(mydata, Inspection.Date >= "2006-01-01" & Inspection.Date <= "2006-12-31")

> df2019 <- subset(mydata, Inspection.Date >= "2019-01-01" & Inspection.Date <= "2019-12-31")

> mean(df2006$Inspection.Score, na.rm = TRUE)

[1] 12.92155

> mean(df2019$Inspection.Score, na.rm = TRUE)

[1] 12.19552

> sd(df2006$Inspection.Score, na.rm = TRUE)
```

```
[1] 18.82576

> sd(df2019$Inspection.Score, na.rm = TRUE)

[1] 17.86964

> df2020 <- subset(mydata, Inspection.Date >= "2020-01-01" & Inspection.Date <=
"2020-12-31")
```