02.21.20 Meeting w/Justine

Panel 15

DataDog: Okay – it looks like we now have two files on my portable, transferred from that ancient Univac tape. The first is carbuncle\_phosdex.csv, the second carbuncle\_lives\_distance.csv. Within Rstudio, I first set my working directory to the location of the files, using the setwd() function. Then I’ll import both files with the read.csv() function.

Panel 16

setwd()

read.csv()

Panel 17

DataDog: Digging deeper into these two files… the dim(), names(), and str() functions will help me figure out what’s in these two dataframes.

Panel 18

DataDog: But what does the data actually look like? I’ll enter the dataframe name by itself in order to see the entire dataset. Let me start with the carbuncle\_lives\_distance\_df.

New Panel

Carbuncle\_lives\_distance\_df

Panel 19

Whoah Kitty! Over 4,000 observations in the carbuncle\_lives\_distance dataframe! Too much. Let’s see just the first few observations. I can use the head() function to look at the first 6 rows.

Panel 20

… and the tail() function to see the last 6 rows.

Panel 21

DataDog: … and now for the carbuncle\_phosdex\_df. I bet this data set contains the phosdex concentrations at various locations around Florida. I wonder how StatCat’s family collected all this data? Hears StatCat snoring… We’ll just need to repeat the same steps for this file. I think I’ll let you guys out there handle that…

Panel 22

Student Exercise…

Panel 23, part B

Listen Mr. Dial-up, there are more than 4000 rows in the carbuncle\_lives\_distance dataframe and more than 400 rows in the carbuncle\_phosdex dataframe.

Panel 24, part B

StatCat: Did I hear somebody say plan?

Panel 25

StatCat in front of projector with the 5 analysis steps listed. Repeat panel 40.

1. Define your research questions
2. Identify the datasets to be used
3. Set inclusion/exclusion criteria
4. Select appropriate statistical tests
5. Interpret results

A little dialogue… this is where the mouse hits the floor.

New Panel

I must answer two research questions in order to defend my illustrious antecedents.

Question the first – What is the relationship between phosdex concentration and distance from the Fountain of youth.

Question the second – What is the relationship between phosdex concentration and cat lives?

Panel 26, part A

StatCat: My little furry friend … now that you and your student helpers have staged the two data files. I wish to share with you how all this data was collected…

DataDog (thinking): Can he read my mind?

This panel is devoted to describing the data collection for the carbuncle\_phosdex data

Panel 27

StatCat: And with regards to the carbuncle\_lives\_distance data, my illustrious great uncle Carbuncle spent weeks inputting this data into the first rolltop UNIVAC. Mystery regarding the provenance of this dataset.

New Panel

StatCat: You will see that both data files share a common field called geoid…

Panel 29 AS-IS

Panel 30

DataDog: What’s that mean StatCat?

StatCat: Without a proper plan, we’ll flounder like Tuna. In this case, we can skip step 3 of our analysis plan because we’re gonna look at all of the data. However, we MUST determine the statistical test we will employ. Because we’re using continuous variables, a simple correlation test will do, resulting in a correlation coefficient.

Panel 32

DataDog: I’m gonna use the merge() function to glue the carbuncle\_phosdex and carbuncle\_lives\_distance dataframes together, resulting in a single dataframe.

Panel 40

StatCat: The data is in. We have now arrived at the penultimate moment, the final step of our analysis plan where our results are interpreted.

DataDog: What’s that mean, StatCat, “the data is in?”

Panel 41

StatCat: My hypothesis is that there is a relationship between the number of cat lives lived and the concentration of phosdex in the water. This WILL BE a positive correlation. The higher the phosdex concentration, the greater number of lives a cat will live.

But to make this point, I must digress and explain positive, negative, and zero correlations.

Panel 43

StatCat: We will not see a zero correlation with my uncle Carbuncle’s data!

StatCat (Commanding tone): My little furry friend, show me the graphs.

DataDog: Okay – but I’ve got to write a couple lines of code… I’ll use the plot() function here!

Damn Code Box

Panel 44 – Revised, reactions, process, etc…

DataDog show the concentration vs. distance plot first – this is negative and highly correlated.

DataDog then shows the concentration vs. age plot – this is positive and highly correlated.

StatCat goes into a CATNIP FRENZY when he sees the second plot.