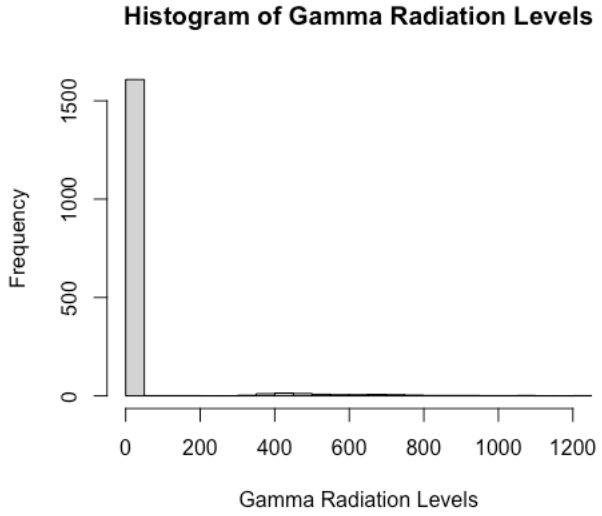
Name: Astrid Fuentes
Date: February 26th, 2021
Title: Final Project – Part II

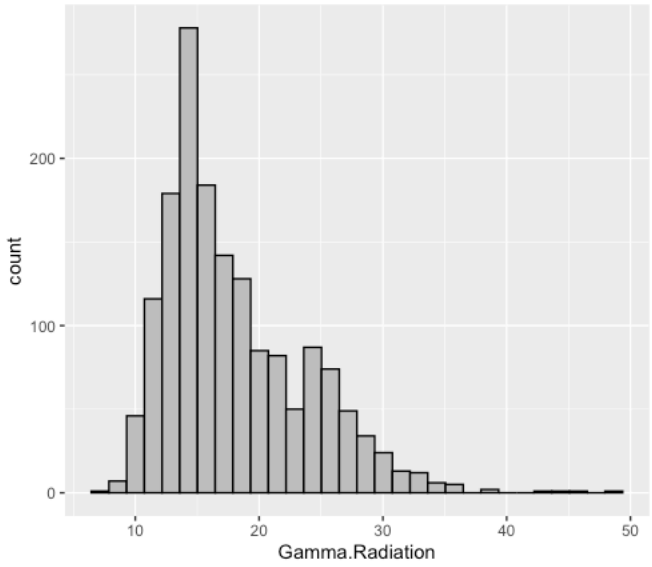
- **Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.**

I imported and examined my data set. Initially I had 7 variables and 1912 rows. I found the variables to be named in an easy-to-understand way, so I decided to not rename them. Next, I decided to change some of the strings to factors as I thought appropriate. For example, the Quarter column contains 4 levels for each quarter of the year. I also looked for and dropped NA values in certain columns I consider key for my analysis, for example the Gamma. Radiation variable. I also found a typo in the Location variable where "Capital District" was misspelled as "Captail District" in several columns. These typos were fixed using the replace() function. Initially, I considered having to transform some Gamma.Radiation variables to have them all in the same unit, upon further inspection, I realized all values were reported in the same unit "mrem/quarter" so this transformation was not necessary.
As far as outliers go, I performed statistics and histogram of the Gamma.Radiation levels which makes me suspect the presence of outliers. The minimum gamma radiation value is 7, the mean is 47, however the max goes all the way up to 1208 while the 3rd quartile remains low at 23. The histogram confirms there are a small number of value that are suspected outliers with radiation levels above 200 mrem/quarter.

```
> summary(data$Gamma.Radiation)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
7.00   14.00   16.90   47.27   23.00 1208.00
```

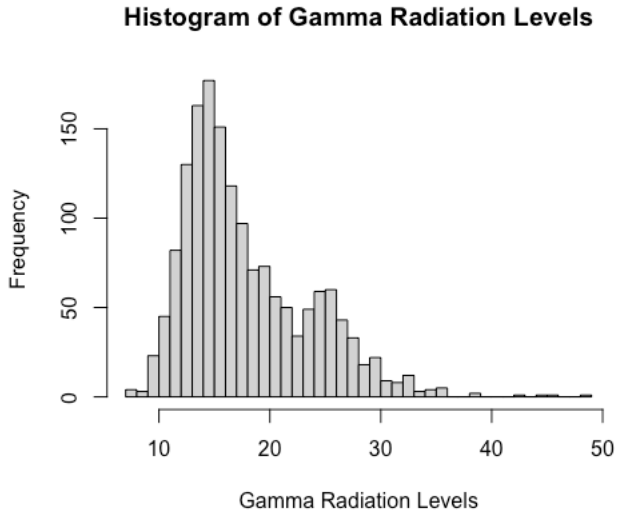


Upon inspecting the comment variable, I found that the West Valley Site Facility in the Fence Line at Waste Area location had "Elevated radiation levels are expected at this monitoring location due to its proximity to a High Level Radioactive Waste Storage Facility. This is an on-site location. Access to this location is controlled by the site operator." I decided to perform a different histogram using ggplot to only show values of Gamma.Radiation <= 50 mrem/quarter. This shows a better distribution of my data:



- **With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.**

The histogram of Gamma Radiation Levels in the cleaned data set suggest the data could be bi-modal with one mode around 15 and another around 25 mrem/quarter. We can still see a few values that go to the upper ranges of 40 and 50 mrem/quarter hence the skewness of the histogram.



In the image below we can see different descriptive statistics of the cleaned data set including a summary(). We can see the data goes from the year 1995 to 2018. We have 7 different Facility Operators in 35 Locations with quarterly measurements being reported in 1608 records.

```
> ## Descriptions of the cleaned data set
> str(data2)
'data.frame':   1608 obs. of  7 variables:
 $ Facility.Operator: Factor w/ 7 levels "Background","Cintichem, Tuxedo, Orange County",..: 3 4 7 7 7 3 4 4 4 5 ...
 $ Location         : Factor w/ 35 levels " Broadway & Bleakly",..: 33 3 28 28 13 25 3 1 3 8 ...
 $ Year             : int  2006 2010 2010 1997 2004 2003 1995 2010 2008 2001 ...
 $ Quarter          : Factor w/ 4 levels "1st","2nd","3rd",..: 4 4 4 2 1 4 4 4 4 4 ...
 $ Gamma.Radiation  : num  21.4 15 15 16 18.1 15.4 13.2 16 20.1 11.8 ...
 $ Reported.Unit    : chr  "mrem/quarter" "mrem/quarter" "mrem/quarter" "mrem/quarter" ...
 $ Comment          : chr  "" "" "" "" ...
> head(data2)
  Facility.Operator                  Location Year Quarter Gamma.Radiation Reported.Unit Comment
1     Ginna Station            Training Center 2006     4th            21.4  mrem/quarter
2      Indian Point                 NYU Tower 2010     4th            15.0  mrem/quarter
3  West Valley Site     Route 240 Zefer's Farm 2010     4th            15.0  mrem/quarter
4  West Valley Site     Route 240 Zefer's Farm 1997     2nd            16.0  mrem/quarter
5  West Valley Site Dutch Hill & Schrartz Road 2004     1st            18.1  mrem/quarter
6     Ginna Station               Parking Lot 2003     4th            15.4  mrem/quarter
> summary(data2)
                    Facility.Operator                          Location         Year      Quarter   Gamma.Radiation
 Background                     : 84     County Route 29 & Miner Road:  86   Min.   :1995   1st:413   Min.   : 7.00
 Cintichem, Tuxedo, Orange County: 48     Training Center             :  86   1st Qu.:1999   2nd:403   1st Qu.:13.90
 Ginna Station                  :256     Webster Sub-Station         :  86   Median :2004   3rd:392   Median :16.40
 Indian Point                   :226     Lakeview Road               :  85   Mean   :2005   4th:400   Mean   :18.04
 Nine Mile Point Site           :339     Capital District            :  84   3rd Qu.:2009             3rd Qu.:21.40
 Shoreham, Suffolk County       : 16     Parking Lot                 :  84   Max.   :2018             Max.   :48.50
 West Valley Site               :639     (Other)                     :1097
 Reported.Unit        Comment
 Length:1608        Length:1608
 Class :character   Class :character
 Mode  :character   Mode  :character

> nrow(data2)
[1] 1608
> ncol(data2)
[1] 7
> head(data2)
  Facility.Operator                  Location Year Quarter Gamma.Radiation Reported.Unit Comment
1     Ginna Station            Training Center 2006     4th            21.4  mrem/quarter
2      Indian Point                 NYU Tower 2010     4th            15.0  mrem/quarter
3  West Valley Site     Route 240 Zefer's Farm 2010     4th            15.0  mrem/quarter
4  West Valley Site     Route 240 Zefer's Farm 1997     2nd            16.0  mrem/quarter
5  West Valley Site Dutch Hill & Schrartz Road 2004     1st            18.1  mrem/quarter
6     Ginna Station               Parking Lot 2003     4th            15.4  mrem/quarter
```

- **What do you not know how to do right now that you need to learn to import and cleanup your dataset?**

  Up to this point, I was able to import my data and do the cleanup process I needed to do. There were a few things I did not remember how to do that I have to look up from previous assignments, for example how to do a filtered histogram using ggplot and how to use the replace () function.

- **Discuss how you plan to uncover new information in the data that is not self-evident.**

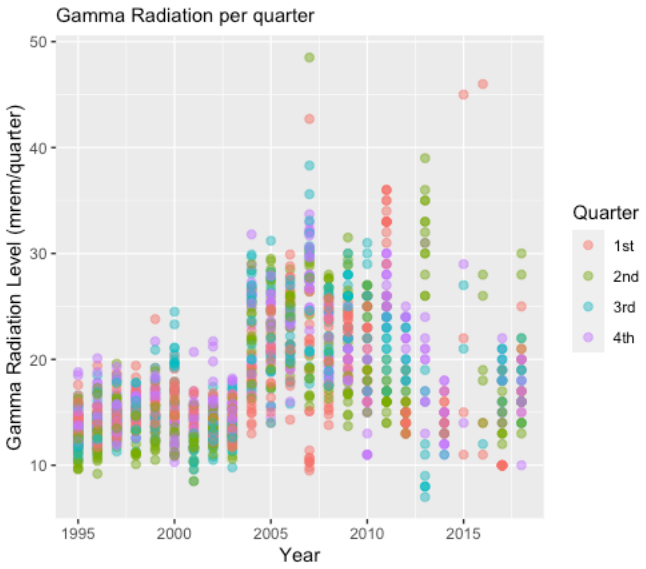
  I would like group the data by Facility/Operator and look at their individual distribution of radiation levels. This would help me uncover additional information regarding the areas that have larger radiation levels.

  I am also interested in performing multiple regression analysis to help me predict radiation levels for the different facilities.

  I also want to do plots over time to see any trends in the radiation levels. I am interested in discovering if these have been increasing over time.

- **What are different ways you could look at this data to answer the questions you want to answer?**

I think some visualizations like line plots and scatter plots will definitely help me answer some of my questions. Additionally, I will need to build my regression model and evaluate if it is a good fit to be able to predict the radiation levels for future years > 2018.



- **Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.**
  As part of my data cleaning process, I already dropped radiation levels > 50. However, I might be interested in saving these in a separate data frame for further analysis. These values could be used to build a regression model for this particular Facility/Operator given its conditions.
  I am considering slicing the data given my suspicious of a bimodal distribution. I will need to review this section of the book one more time to find the best way to deal with this.
  I do not expect to create new variables or join a separate data frame to be able to answer my questions.

- **How could you summarize your data to answer key questions?**
  In my opinion, visualizations are worth 1000 words. I am confident that plotting my data correctly will be able to give me great insights that I can then explain and summarize in written.

- **What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).**
  I will be performing scatter plots and line plots in addition to the histograms I showed above. Besides summary tables and regression outputs, the only other table I consider using is a correlation table.

  Using the cor.test function we can see that the Gamma.Radiation values are positively correlated with the Year r=0.4 with a significant p-value < 2.2e-16.

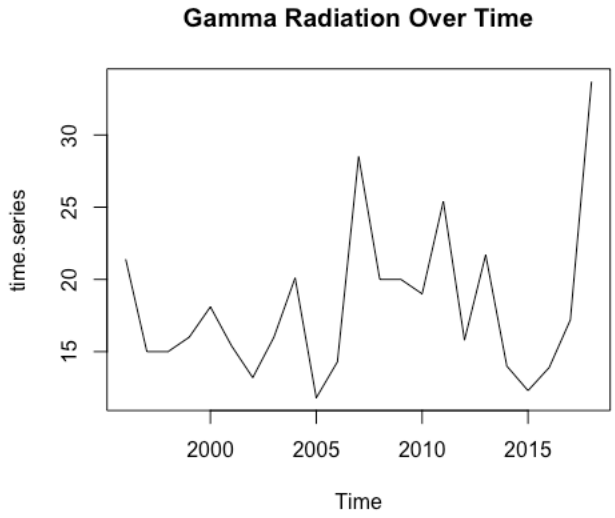
  > cor.test(data2$Gamma.Radiation, data2$Year, method=c("pearson", "kendall", "spearman"))

  Pearson's product-moment correlation

  data: data2$Gamma.Radiation and data2$Year
  t = 17.662, df = 1606, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3615381 0.4434359
sample estimates:
      cor
0.4032943

I was able to build a time series of Gamma Radiation levels over time showing. There seem to be a couple of spikes around 2004, 2007, 2011, 2013, and a large spike towards 2018 as shown below:



- **What do you not know how to do right now that you need to learn to answer your questions?**
  I know I have done this before, but I can't build some of the plot commands from the top of my head. I will be referring to my previous assignments in order to get all my plots done correctly including the labels, colors, etc.
  I would like to do a matrix scatter plot of Gamma.Radiation per year separated by groups of quarters. I think this would help see if there are quarters where radiation is consistently larger or smaller than the rest. I am not sure how this is done so I will need to keep trying.
  I would like to further explore the time series to see if I am able to make similar plots per quarter and per Facility/Operator.

- **Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**
  I can try to apply some clustering techniques to this data set. I will need to look into it a little further.