Are Volcanic Eruptions Increasing?: An Example of Teaching Data Wrangling and Visualization in Stat 2

Kelly McConville, Swarthmore College

Wednesday, May 18th

Data Science in Stat 2?

- **ASA's Curriculum Guidelines for Undergraduate Programs in Statistical Science**
  - "Increased importance of data science"
  - "the ability to access and manipulate data in various ways"
  - "The statistical analysis process involves … considering whether available data are appropriate for addressing the problem, … undertaking analysis in a reproducible manner"
  - "For programs that are unable to implement an entire major program, we suggest that missing topics or skills be added to classes in the current curriculum."

- I want to give you one example of how to add data wrangling and visualization into Stat 2.

Data and Software

- **Smithsonian Global Volcanism Program** maintains two databases:
  - **Volcanoes of the World**
  - **Eruptions**
- RStudio
  - **R Markdown**
- R Packages
  - **ggplot2**
  - **dplyr**

Are Volcanic Eruptions Increasing?

```
Eruptions
```

```
Source: local data frame [10,770 x 22]

   Volcano.Number  Volcano.Name Eruption.Number      Eruption.Type
            (int)        (fctr)            (int)             (fctr)
1          311240     Cleveland            20845 Confirmed Eruption
2          263250        Merapi            20842 Confirmed Eruption
3          261080      Sinabung            20818 Confirmed Eruption
4          211060          Etna            20841 Confirmed Eruption
5          300260  Kliuchevskoi            20786 Confirmed Eruption
6          290200         Ketoi            20780 Uncertain Eruption
7          357120    Villarrica            20838 Confirmed Eruption
8          252120        Ulawun            20835 Confirmed Eruption
9          284090 Izu-Torishima            20828 Uncertain Eruption
10         354020        Ubinas            20832 Confirmed Eruption
..            ...           ...              ...               ...
Variables not shown: Area.of.Activity (fctr), VEI (int), VEI.Modifier
  (fctr), Start.Year.Modifier (fctr), Start.Year (int),
  Start.Year.Uncertainty (int), Start.Month (int), Start.Day.Modifier
  (fctr), Start.Day (int), Start.Day.Uncertainty (int),
  Evidence.Method..dating. (fctr), End.Year.Modifier (fctr), End.Year
  (int), End.Year.Uncertainty (int), End.Month (int), End.Day.Modifier
  (fctr), End.Day (int), End.Day.Uncertainty (int)
```

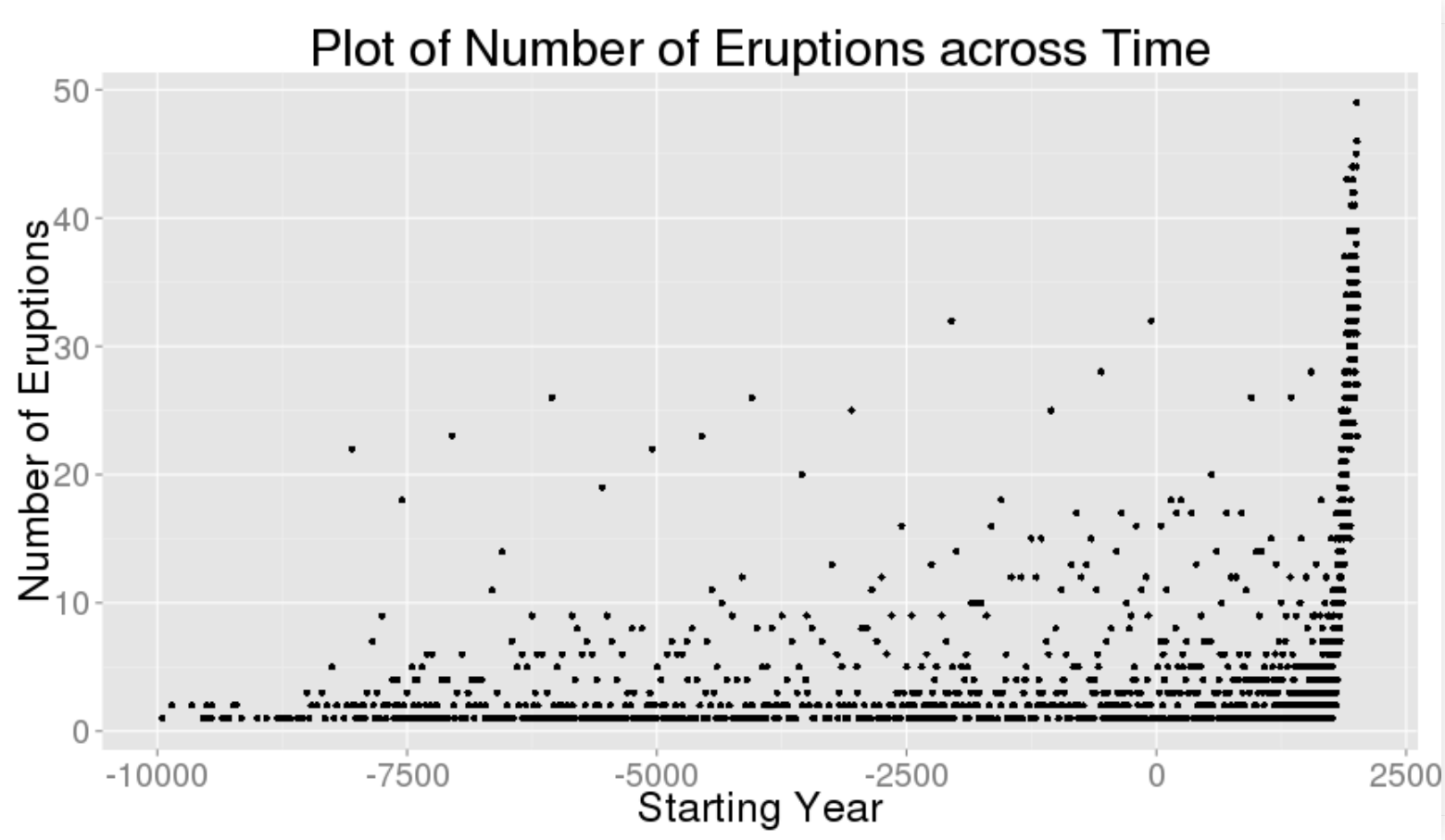
Are Volcanic Eruptions Increasing?

- Need to wrangle the data
  - Subset to include only confirmed eruptions
  - Want each row to represent a year, not an eruption

```
Eruptions.Year <-
  Eruptions %>%
  subset(Eruption.Type=="Confirmed Eruption") %>%
  group_by(Start.Year) %>%
  summarise(count = length(Start.Year), avg.VEI= mean(VEI, na.rm=TRUE))
Eruptions.Year
```

```
Source: local data frame [1,501 x 3]

   Start.Year count avg.VEI
        (int) (int)   (dbl)
1       -9950     1     NaN
2       -9850     2     NaN
3       -9650     2       5
4       -9540     1     NaN
5       -9520     1     NaN
6       -9500     2       2
7       -9490     1       3
8       -9460     1       5
9       -9450     2       5
10      -9350     1       5
..        ...   ...     ...
```

Are Volcanic Eruptions Increasing?



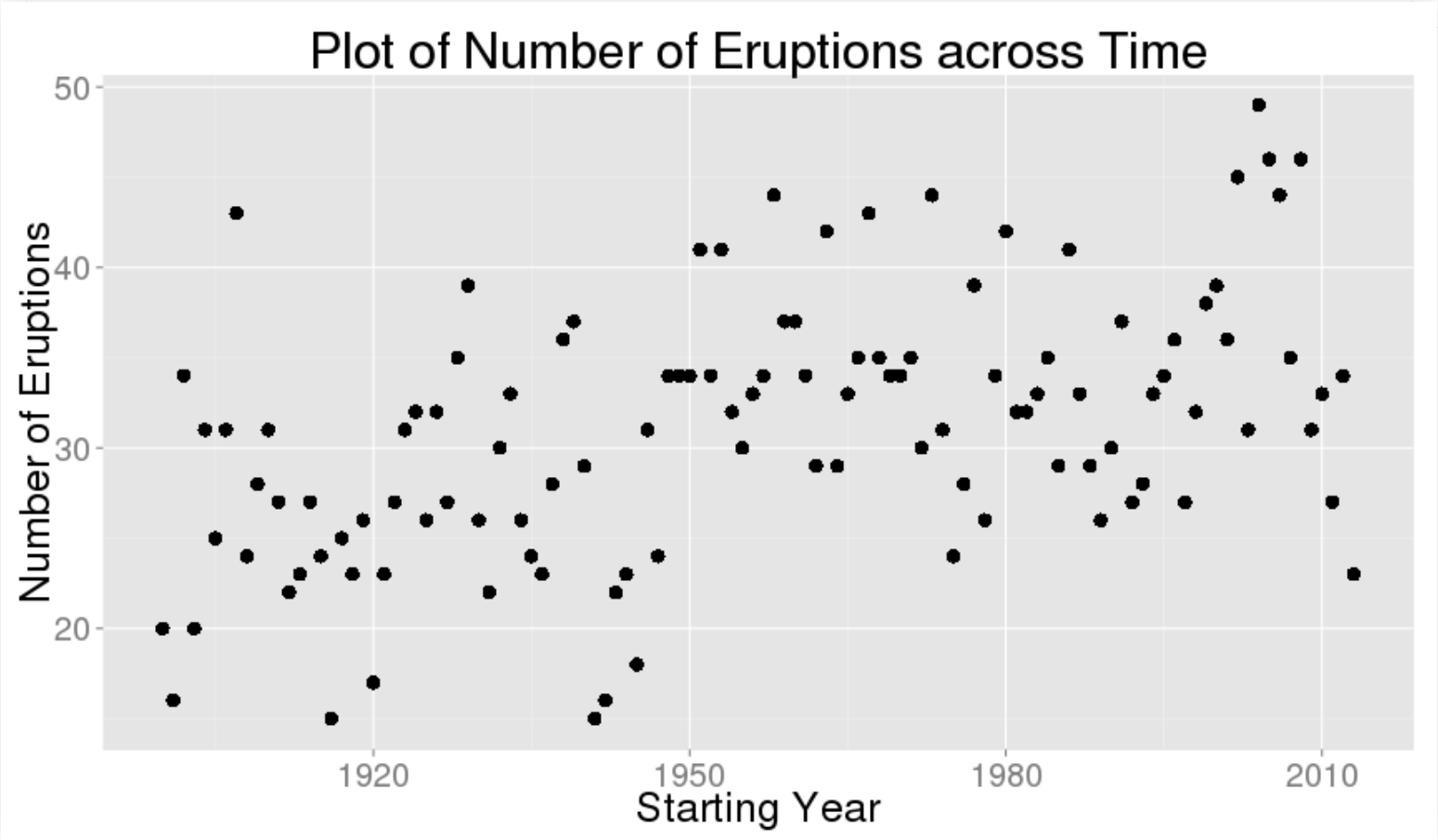
Are Volcanic Eruptions Increasing?

- Focus on 1900 to the present

```
Eruptions.Year <-
  Eruptions %>%
  filter(Start.Year>=1900) %>%
  subset(Eruption.Type=="Confirmed Eruption") %>%
  group_by(Start.Year) %>%
  summarise(count = length(Start.Year), avg.VEI= mean(VEI, na.rm=TRUE))
Eruptions.Year
```

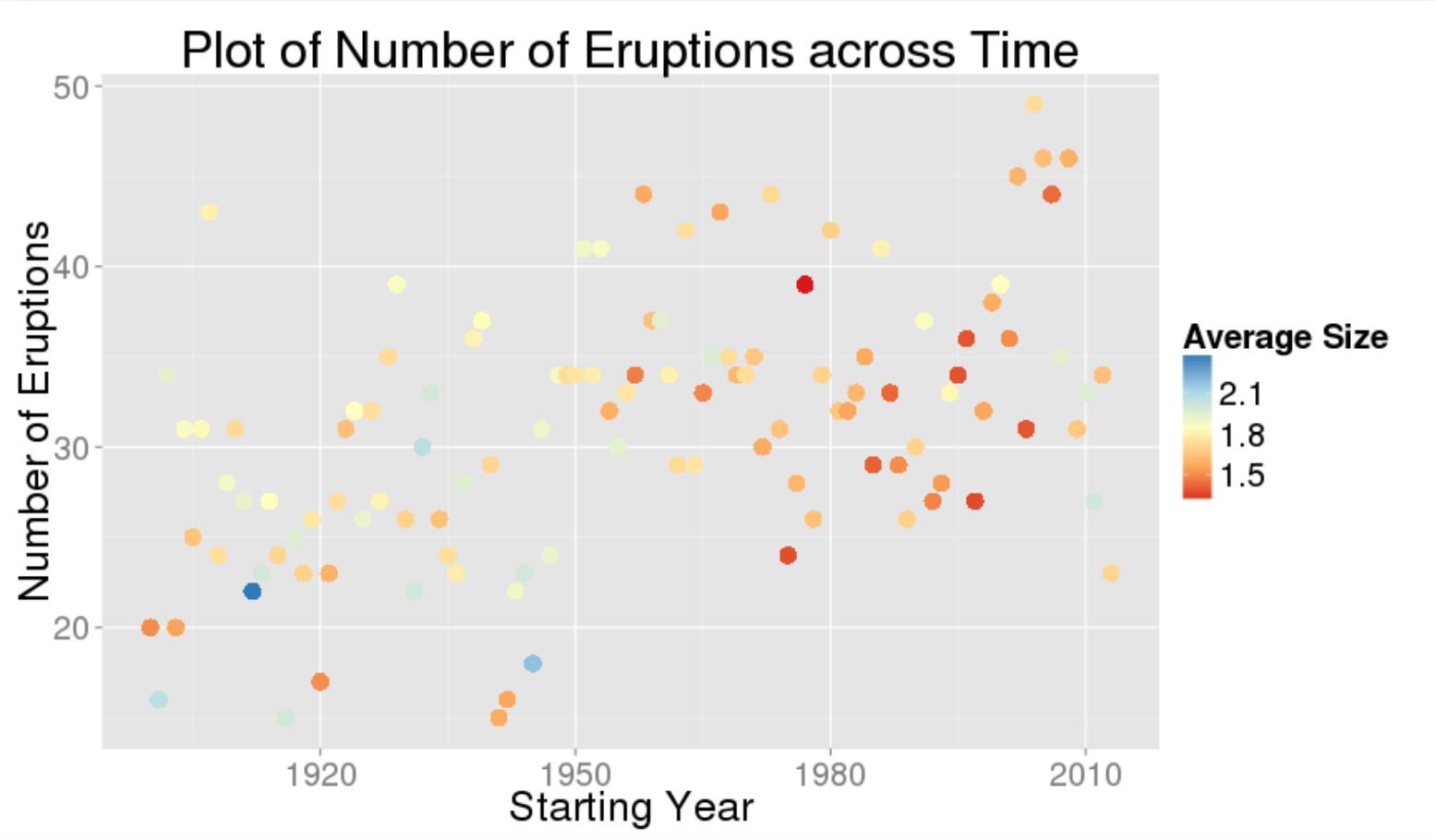
```
Source: local data frame [114 x 3]

   Start.Year count  avg.VEI
        (int) (int)    (dbl)
1        1900    20 1.500000
2        1901    16 2.066667
3        1902    34 1.941176
4        1903    20 1.555556
5        1904    31 1.866667
6        1905    25 1.652174
7        1906    31 1.827586
8        1907    43 1.800000
9        1908    24 1.739130
10       1909    28 1.888889
..        ...   ...      ...
```

Are Volcanic Eruptions Increasing?



- Sampling Bias?

Are Volcanic Eruptions Increasing?



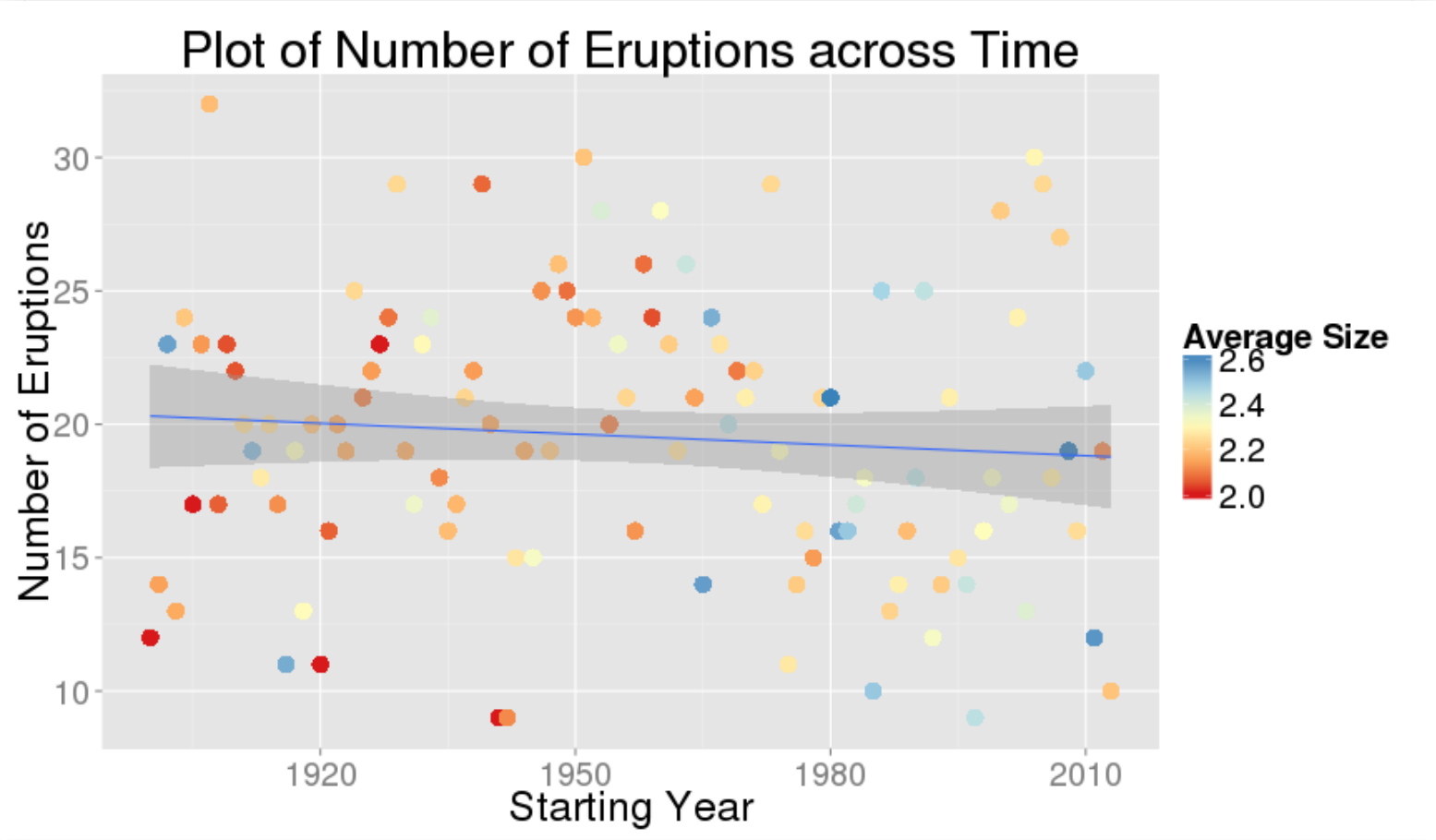
Are Volcanic Eruptions Increasing?

- Additional filter: size

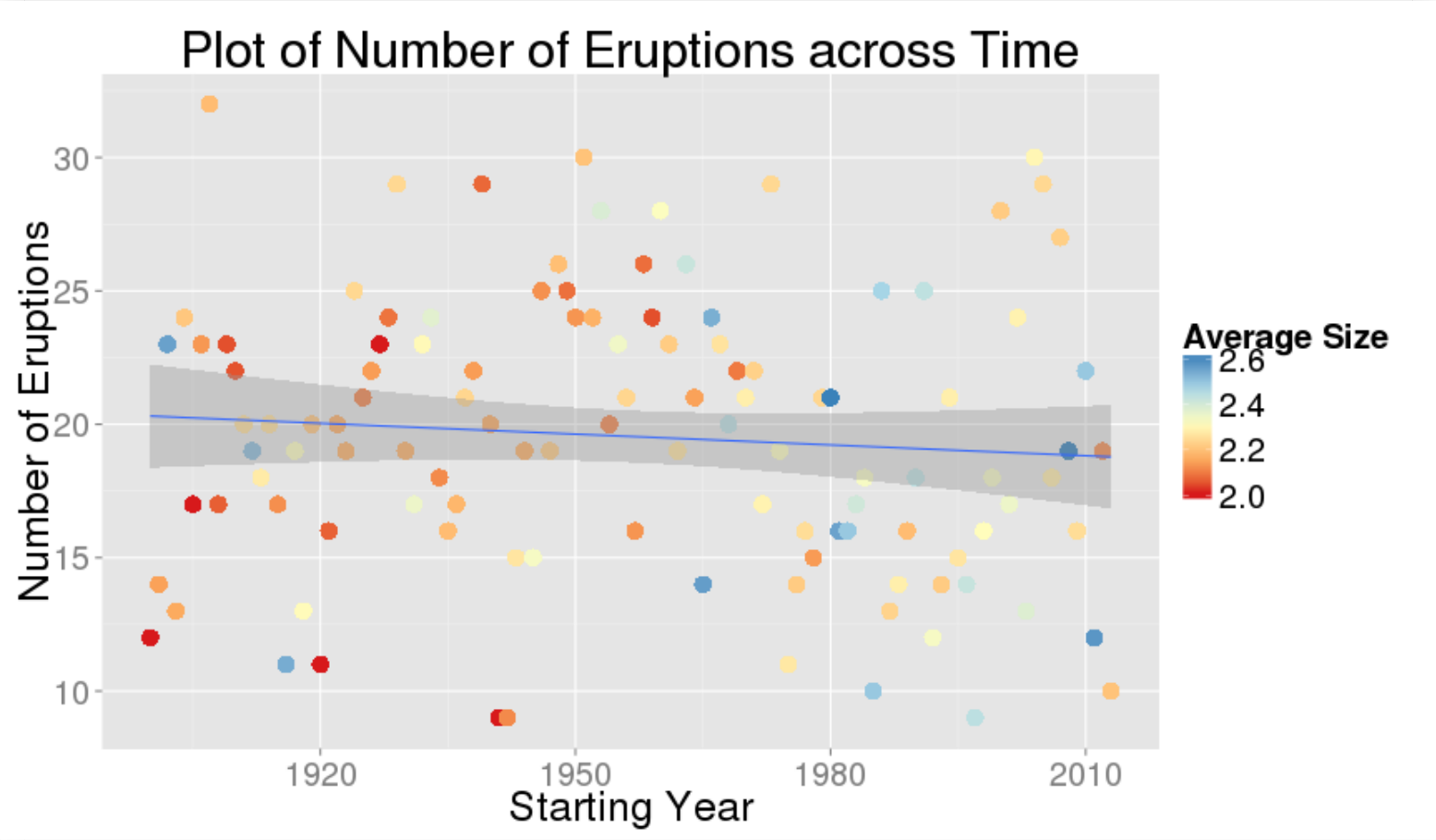
```
Eruptions.Year <-
  Eruptions %>%
  filter(Start.Year>=1900, VEI>=2) %>%
  subset(Eruption.Type=="Confirmed Eruption") %>%
  group_by(Start.Year) %>%
  summarise(count = length(Start.Year), avg.VEI= mean(VEI, na.rm=TRUE))
Eruptions.Year
```

```
Source: local data frame [114 x 3]

   Start.Year count  avg.VEI
        (int) (int)    (dbl)
1        1900    12 2.000000
2        1901    14 2.142857
3        1902    23 2.565217
4        1903    13 2.153846
5        1904    24 2.208333
6        1905    17 2.000000
7        1906    23 2.130435
8        1907    32 2.187500
9        1908    17 2.058824
10       1909    23 2.043478
..        ...   ...      ...
```

Are Volcanic Eruptions Increasing?



Are Volcanic Eruptions Increasing?



Other Questions

- Duration
  - Have to manipulate start and end dates!
- Duration and Size
- Location
  - Merge two datasets
- Interactive Visualizations
  - **Map of World's Volcanoes**

References

- J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, A. Atkins and R. Hyndman (2015). rmarkdown: Dynamic Documents for R. R package version 0.8. http://CRAN.R-project.org/package=rmarkdown

- American Statistical Association Undergraduate Guidelines Workgroup. 2014. 2014 curriculum guidelines for undergraduate programs in statistical science. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/curriculumguidelines.cfm

Global Volcanism Program, 2013. Volcanoes of the World, v. 4.4.3. Venzke, E (ed.). Smithsonian Institution. Downloaded 06 May 2016. http://dx.doi.org/10.5479/si.GVP.VOTW4-2013

References

- H. Wickham and R. Francois (2015). dplyr: A Grammar of Data Manipulation. R package version 0.4.3. **http://CRAN.R-project.org/package=dplyr**

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.