ClementiPrj1DataExploration

Johnathan Clementi

2/16/2020

Library load

library(tidyverse) #loads ggplot and the others  
library(readxl)#for reading in excel files

Read data - fill blanks with NA

czenSites <- read.csv('C:/Users/clemenj/Google Drive/Grad School/Juniata\_DataScience/DS500/Project1/data/longtermMonitoringSites.csv', na.strings = c("", "NA"))

Metadata:

Origin - CZEN (Critical Zone Exloration Network) / CZO (Critical Zone Observatory) Site Seeker website: <https://www.czen.org/site_seeker>

Date of Creation - Data exported from czen site seeker on July 25, 2018.

List of Variables: 1. Field Site - (Categorical) - Name given to field site 2. Latitude - (Quantitative) - Latitude of field site 3. Longitude - (Quantitative) - Longitude of field site 4. Field Site type - (Categorical) - Affiliation of field site with US or international based research groups 5. Registration - (Categorical) -Is the field site a registered member of CZEN 6. Mean Annual Precipitation - (Quantitative) - Average precipitation in millimiters per year 7. Average Annual Temperature - (Quantitative) - Average temperature in degrees Centigrade 8. Geology - (Categorical) - Most common geology type present at field site 9. Land Cover - (Categorical) - Most common land cover type present at field site 10. Climate - (Categorical) - Most common climate type present at field site 11. Sub-Climate - (Categorical) - Second most common climate type present at field site 12. Network(s) - (Categorical) - Research networks the field site is affiliated with 13. Soil Order - (Categorical) - Most common soil type present at field site 14. Study Start Date - (Quantitative) - The year in which the field site was opened for research 15. Hydrology: Surface water stream order - (Categorical) - Stream order of primary surface water present at field site 16. Hydrology: Surface water - Stream Flow Performance - (Categorical) - Stream flow type of primary surface water present at field site 17. Hydrology: Groundwater - (Categorical) - Type of groundwater present at field site

Data Cleaning and Manipulation

Rename variable headings to exclude spaces

czenSites = czenSites %>%  
 rename(FieldSite = 'Field.Site', FieldSiteType = 'Field.Site.Type', MeanAnnualPrecip = 'Mean.Annual.Precipitation', AvgTemp = 'Average.Annual.Temperature', LandCover = 'Land.Cover', SubClimate = 'Sub.Climate', Networks = 'Network.s.', SoilOrder = 'Soil.Order', StudyStartDate = 'Study.Start.Date', SurfaceStreamOrder = 'Hydrology..Surface.water.stream.order', StreamFlowPerformance = 'Hydrology..Surface.water...Stream.Flow.Performance', Groundwater = 'Hydrology..Groundwater')  
  
str(czenSites)

## 'data.frame': 233 obs. of 17 variables:  
## $ FieldSite : Factor w/ 233 levels "Abby Road","Abisko Research Station",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Latitude : num 45.8 68.3 43.8 47.9 -22.5 ...  
## $ Longitude : num -122.33 18.82 -74.84 -4.15 133.34 ...  
## $ FieldSiteType : Factor w/ 9 levels "Independent",..: 9 8 1 3 3 9 3 9 9 9 ...  
## $ Registration : Factor w/ 2 levels "Registered","Unregistered": 2 2 1 2 2 1 1 2 2 2 ...  
## $ MeanAnnualPrecip : Factor w/ 158 levels "100 millimeters / year",..: NA 89 17 15 88 32 118 66 119 86 ...  
## $ AvgTemp : Factor w/ 36 levels "-12¡C","-2¡C",..: NA 9 30 11 24 29 26 35 23 NA ...  
## $ Geology : Factor w/ 66 levels "carbonate","carbonate, igneous-felsic intrusive",..: NA 15 9 9 NA 25 NA 24 1 47 ...  
## $ LandCover : Factor w/ 92 levels "Agriculture- Crops",..: 42 22 90 4 60 90 24 39 89 83 ...  
## $ Climate : Factor w/ 5 levels "Boreal","Mediterranean",..: NA 3 4 4 NA 4 5 4 5 3 ...  
## $ SubClimate : Factor w/ 9 levels "Arid","Dry","Humid",..: NA NA 7 3 NA 7 6 4 7 NA ...  
## $ Networks : Factor w/ 58 levels "ARS","ARS, CUAHSI",..: 46 52 32 14 56 57 51 38 35 35 ...  
## $ SoilOrder : Factor w/ 52 levels "Alfisol","Alfisol, Aridisol, Inceptisol, Mollisol",..: NA 34 49 51 51 51 NA 12 26 34 ...  
## $ StudyStartDate : Factor w/ 92 levels "1907","1910",..: 92 4 74 NA 90 89 59 22 17 92 ...  
## $ SurfaceStreamOrder : Factor w/ 12 levels "Fifth Order",..: NA NA 2 NA NA 2 4 10 NA NA ...  
## $ StreamFlowPerformance: Factor w/ 7 levels "Ephemeral","Intermittent",..: NA 4 2 NA NA 4 3 4 NA NA ...  
## $ Groundwater : Factor w/ 5 levels "Arrangment of Aquifer Components - Complexly interbedded sequence of aquifers and confining beds",..: NA NA NA 2 NA 2 2 NA 1 NA ...

Fix misspelled values in FieldSiteType, Geology, & Landcover

czenSites = czenSites %>%  
 mutate(FieldSiteType = str\_replace(FieldSiteType, "Interntaional", "International")) %>%  
 mutate(Geology = str\_replace(Geology, "secimentary", "sedimentary")) %>%  
 mutate(LandCover = str\_replace(LandCover, "Shurbland", "Shrubland"))

Replace Unknown / None values with NA in Networks, SoilOrder, and SurfaceStreamOrder

czenSites = czenSites %>%  
 mutate(Networks = str\_replace(Networks, "None", NA\_character\_)) %>%  
 mutate(SoilOrder = str\_replace(SoilOrder, "Unknown", NA\_character\_)) %>%  
 mutate(SurfaceStreamOrder = str\_replace(SurfaceStreamOrder, "Unknown", NA\_character\_))

Remove units from MeanAnnualPrecip, then convert to integer

#Use mutate to change the existing data in the variable, use str\_replace (part of stringr) and regex to identify unit string and replace it with empty string  
  
czenSites = czenSites %>%  
 mutate(MeanAnnualPrecip = str\_replace(MeanAnnualPrecip, "millimeters / year", "")) %>%  
 mutate(MeanAnnualPrecip = as.numeric(MeanAnnualPrecip))

## Warning: NAs introduced by coercion

Remove units from AvgTemp, then convert to integer & remove dubious value (73C = 163F & 34C = 93.2F) from AvgTemp

czenSites = czenSites %>%  
 mutate(AvgTemp = str\_replace(AvgTemp, "¡C", "")) %>%  
 mutate(AvgTemp = str\_replace(AvgTemp, "73", NA\_character\_)) %>%  
 mutate(AvgTemp = str\_replace(AvgTemp, "34", NA\_character\_)) %>%  
 mutate(AvgTemp = as.numeric(AvgTemp))

Remove date ranges for StudyStartDate and convert to integer

czenSites = czenSites %>%  
 mutate(StudyStartDate = str\_replace(StudyStartDate, " to .\*", "")) %>%  
 mutate(StudyStartDate = as.numeric(StudyStartDate))

Remove redundant text from Groundwater values

czenSites = czenSites %>%  
 mutate(Groundwater = str\_replace(Groundwater, "Arrangment of Aquifer Components - ", "")) %>%  
 mutate(Groundwater = str\_replace(Groundwater, "Arrangment of Aquifer Components - ", ""))

Simplify Groundwater Values

czenSites = czenSites %>%  
 mutate(Groundwater = str\_replace(Groundwater, "Single, dominant unconfined aquifer", "Single unconfined aquifer")) %>%  
 mutate(Groundwater = str\_replace(Groundwater, "Complexly interbedded sequence of aquifers and confining beds", "Combination of aquifers and confining beds")) %>%  
 mutate(Groundwater = str\_replace(Groundwater, "Three unit system of an unconfined aquifer, confined aquifer and a confining bed", "Unconfined and confined aquifer and confining bed"))

Split multivalue observations in variables for Geology, Landcover, SubClimate, Networks, SoilOrder, SurfaceStreamOrder, StreamFlowPerformance, Groundwater. Then select relevant variables.

* Note - These functions create new dataframes to preserve the count for only that variable (If I did these operations on the same dataframe, values would be counted more than once for observations that have multiple multi-part values.)

# Geology  
czenSitesGeology = czenSites %>%  
 transform(Geology = strsplit(as.character(Geology), ",")) %>%  
 unnest(Geology) %>%  
 mutate(Geology = str\_trim(Geology)) %>%  
 select(FieldSite, Latitude, Longitude, Geology)  
  
# LandCover  
czenSitesLandCover = czenSites %>%  
 transform(LandCover = strsplit(as.character(LandCover), ",")) %>%  
 unnest(LandCover) %>%  
 mutate(LandCover = str\_trim(LandCover)) %>%  
 select(FieldSite, Latitude, Longitude, LandCover)  
  
# SubClimate   
czenSitesSubClimate = czenSites %>%  
 transform(SubClimate = strsplit(as.character(SubClimate), ",")) %>%  
 unnest(SubClimate) %>%  
 mutate(SubClimate = str\_trim(SubClimate)) %>%  
 select(FieldSite, Latitude, Longitude, Climate, SubClimate)  
  
# Networks  
czenSitesNetworks = czenSites %>%  
 transform(Networks = strsplit(as.character(Networks), ",")) %>%  
 unnest(Networks) %>%  
 mutate(Networks = str\_trim(Networks)) %>%  
 select(FieldSite, Latitude, Longitude, Networks)  
  
# Soil Order  
czenSitesSoilOrder = czenSites %>%  
 transform(SoilOrder = strsplit(as.character(SoilOrder), ",")) %>%  
 unnest(SoilOrder) %>%  
 mutate(SoilOrder = str\_trim(SoilOrder)) %>%  
 select(FieldSite, Latitude, Longitude, SoilOrder)  
  
# Surface Stream Order  
czenSitesStreamOrder = czenSites %>%  
 transform(SurfaceStreamOrder = strsplit(as.character(SurfaceStreamOrder), ",")) %>%  
 unnest(SurfaceStreamOrder) %>%  
 mutate(SurfaceStreamOrder= str\_trim(SurfaceStreamOrder)) %>%  
 select(FieldSite, Latitude, Longitude, SurfaceStreamOrder)  
  
# Stream Performance  
czenSitesStreamPerformance = czenSites %>%  
 transform(StreamFlowPerformance = strsplit(as.character(StreamFlowPerformance), ",")) %>%  
 unnest(StreamFlowPerformance) %>%  
 mutate(StreamFlowPerformance = str\_trim(StreamFlowPerformance)) %>%  
 select(FieldSite, Latitude, Longitude, StreamFlowPerformance)  
  
# Groundwater  
czenSitesGroundwater = czenSites %>%  
 transform(Groundwater = strsplit(as.character(Groundwater), ",")) %>%  
 unnest(Groundwater) %>%  
 mutate(Groundwater = str\_trim(Groundwater)) %>%  
 select(FieldSite, Latitude, Longitude, Groundwater)

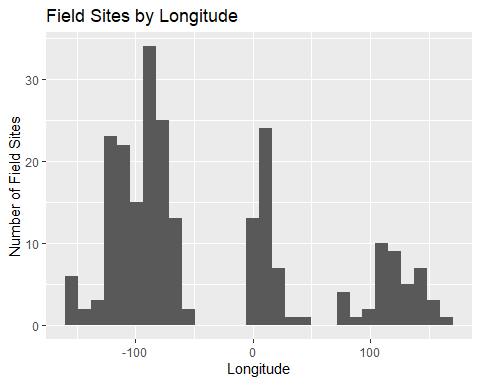
Variation within each Variable

Longitude:

Longitude is the ‘x’ value in a coordinate system used to locate points of interest on the Earth’s surface. Lines of Longitude run from North -> South, but measure East -> West. The Prime Meridian (0 longitude), runs through Greenwich, England. This figure is interesting because you can see that the majority of research sites are located in the US. Then there are not any research sites in the Atlantic Ocean, a few sites in Europe, not many in Africa.

ggplot(data = czenSites) +  
 geom\_histogram(mapping = aes(x = Longitude)) +  
 labs(title = "Field Sites by Longitude", x = 'Longitude', y = "Number of Field Sites")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

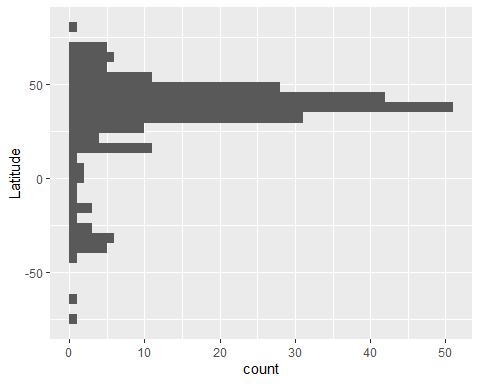


Latitude:

Latitude is the ‘y’ value in a coordinate system used to locate points of interest on the Earth’s surface. Lines of Latitude run from East -> West, but measure North -> South. The Equator (0 latitude), runs through the country of Equador. This figure is interesting because you can observe that most of the research sites are located in the Northern Hemisphere.

ggplot(data = czenSites) +  
 geom\_histogram(mapping = aes(x = Latitude)) +  
 coord\_flip()

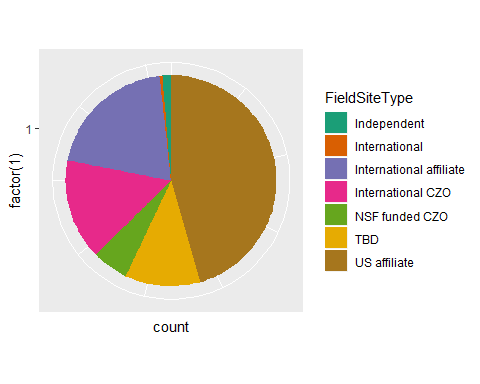
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Type of Field Site:

Traditionally, the US has been a leader in long-term ecological research. This is evident in the number of US-Affiliated research sites, as illustrated by this figure.

ggplot(data = czenSites) +  
 geom\_bar(mapping = aes(x = factor(1), fill = FieldSiteType), width = 1) +  
 coord\_polar("y") +  
 scale\_fill\_brewer(palette = "Dark2") +  
 theme(axis.text.x = element\_blank())



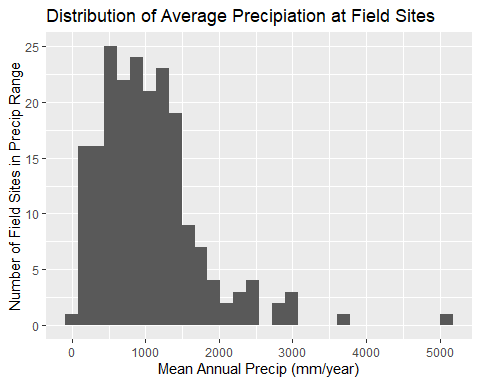
Mean Annual Precip:

One of the goals of this study is to identify ecosystems which lack representation in long-term research studies. In this case, it appears that there may be a lack of long-term research that focuses on areas that would be classified as rainforests (precipitation > 2500 mm/year)

ggplot(data = czenSites) +  
 geom\_histogram(mapping = aes(x = MeanAnnualPrecip)) +  
 labs(title = "Distribution of Average Precipiation at Field Sites", x = "Mean Annual Precip (mm/year)", y = "Number of Field Sites in Precip Range")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 30 rows containing non-finite values (stat\_bin).



summary(czenSites$MeanAnnualPrecip)

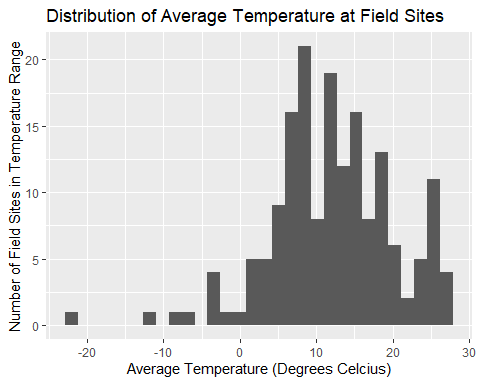
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 50 598 903 1051 1350 5143 30

Average Temperature

ggplot(data = czenSites) +  
 geom\_histogram(mapping = aes(x = AvgTemp)) +  
 labs(title = "Distribution of Average Temperature at Field Sites", x = "Average Temperature (Degrees Celcius)", y = "Number of Field Sites in Temperature Range")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 63 rows containing non-finite values (stat\_bin).



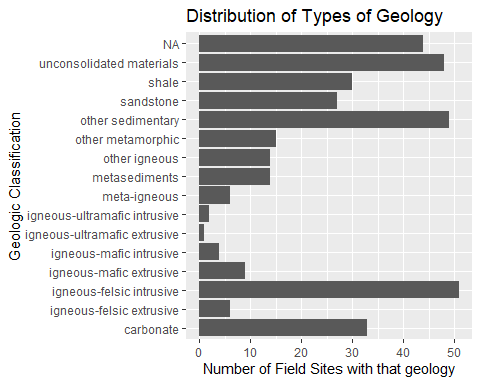
summary(czenSites$AvgTemp)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -22 7 12 12 17 27 63

Geology:

Most geologic types (types of rocks) are represented in the global long-term reserach sites with the exception of igneaous-ultramafic intrusive/extrusive, and igneous-felsic extrusive. These geologic types may be rare, and possibly only exist on small islands or areas that are not conducive to the current long-term research model.

ggplot(data = czenSitesGeology) +  
 geom\_bar(mapping = aes(x = Geology)) +  
 coord\_flip() +  
 labs(title = "Distribution of Types of Geology", y = "Number of Field Sites with that geology", x = "Geologic Classification")



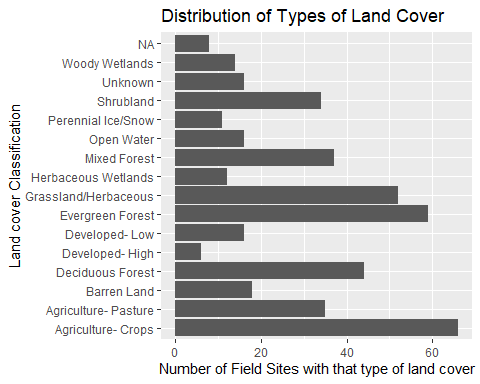
czenSitesGeology %>% group\_by(Geology) %>% count(Geology)

## # A tibble: 16 x 2  
## # Groups: Geology [16]  
## Geology n  
## <chr> <int>  
## 1 carbonate 33  
## 2 igneous-felsic extrusive 6  
## 3 igneous-felsic intrusive 51  
## 4 igneous-mafic extrusive 9  
## 5 igneous-mafic intrusive 4  
## 6 igneous-ultramafic extrusive 1  
## 7 igneous-ultramafic intrusive 2  
## 8 meta-igneous 6  
## 9 metasediments 14  
## 10 other igneous 14  
## 11 other metamorphic 15  
## 12 other sedimentary 49  
## 13 sandstone 27  
## 14 shale 30  
## 15 unconsolidated materials 48  
## 16 <NA> 44

Landcover:

Focusing on less represented research locations, this figure highlights that highly developed areas (cities) are not well represented in research sites as Evergreen forests are. This is consistent with a trend I have observed in ecology, in that scientists are interested in researching relatively unmodified ecosystems. However, there has been a push lately to begin researching urban ecologic systems. For instance, Baltimore has a well known long-term research site.

ggplot(data = czenSitesLandCover) +  
 geom\_bar(mapping = aes(x = LandCover)) +  
 coord\_flip() +  
 labs(title = "Distribution of Types of Land Cover", y = "Number of Field Sites with that type of land cover", x = "Land cover Classification")



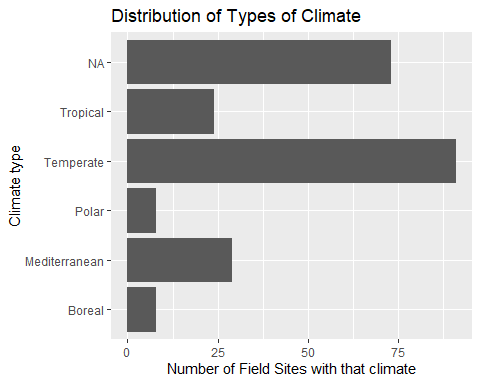
czenSitesLandCover %>% group\_by(LandCover) %>% count(LandCover)

## # A tibble: 16 x 2  
## # Groups: LandCover [16]  
## LandCover n  
## <chr> <int>  
## 1 Agriculture- Crops 66  
## 2 Agriculture- Pasture 35  
## 3 Barren Land 18  
## 4 Deciduous Forest 44  
## 5 Developed- High 6  
## 6 Developed- Low 16  
## 7 Evergreen Forest 59  
## 8 Grassland/Herbaceous 52  
## 9 Herbaceous Wetlands 12  
## 10 Mixed Forest 37  
## 11 Open Water 16  
## 12 Perennial Ice/Snow 11  
## 13 Shrubland 34  
## 14 Unknown 16  
## 15 Woody Wetlands 14  
## 16 <NA> 8

Climate:

Polar and Boreal climates are the least represented across the research sites. This is most likely because these environments are more harsh and less conducive to human living. There are more resources needed to live and conduct research in these locations.

ggplot(data = czenSites) +  
 geom\_bar(mapping = aes(x = Climate)) +  
 coord\_flip()+  
 labs(title = "Distribution of Types of Climate", y = "Number of Field Sites with that climate", x = "Climate type")



czenSites %>% group\_by(Climate) %>% count(Climate)

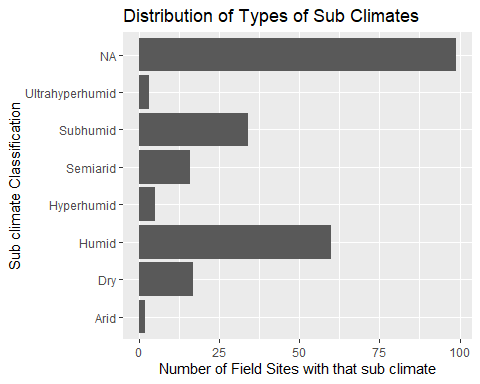
## Warning: Factor `Climate` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`  
  
## Warning: Factor `Climate` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`  
  
## Warning: Factor `Climate` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`  
  
## Warning: Factor `Climate` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 6 x 2  
## # Groups: Climate [6]  
## Climate n  
## <fct> <int>  
## 1 Boreal 8  
## 2 Mediterranean 29  
## 3 Polar 8  
## 4 Temperate 91  
## 5 Tropical 24  
## 6 <NA> 73

Sub Climate:

These results are supportive of those found in the Mean Annual Precipitation figure. Most research sites are in areas that have temperate environments that are not on the extreme ends of humidity. Deserts and rainforests are two areas that are not well represented in long-term research.

ggplot(data = czenSitesSubClimate) +  
 geom\_bar(mapping = aes(x = SubClimate)) +  
 coord\_flip() +  
 labs(title = "Distribution of Types of Sub Climates", y = "Number of Field Sites with that sub climate", x = "Sub climate Classification")



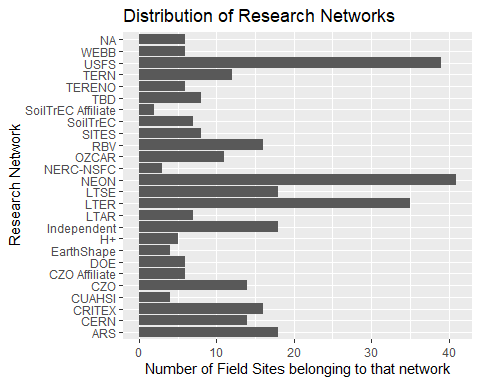
czenSitesSubClimate %>% group\_by(SubClimate) %>% count(SubClimate)

## # A tibble: 8 x 2  
## # Groups: SubClimate [8]  
## SubClimate n  
## <chr> <int>  
## 1 Arid 2  
## 2 Dry 17  
## 3 Humid 60  
## 4 Hyperhumid 5  
## 5 Semiarid 16  
## 6 Subhumid 34  
## 7 Ultrahyperhumid 3  
## 8 <NA> 99

Networks:

A network is essentially an organization of scientists involved in the same type of research. For instance, the USFS network is comprised of US Forest Service members that study forestry in the US. NEON stands for National Ecologic Observation Network and houses scientists that focus ecologic systems rather than geologic systems (CZO).

ggplot(data = czenSitesNetworks) +  
 geom\_bar(mapping = aes(x = Networks)) +  
 coord\_flip() +  
 labs(title = "Distribution of Research Networks", y = "Number of Field Sites belonging to that network", x = " Research Network")



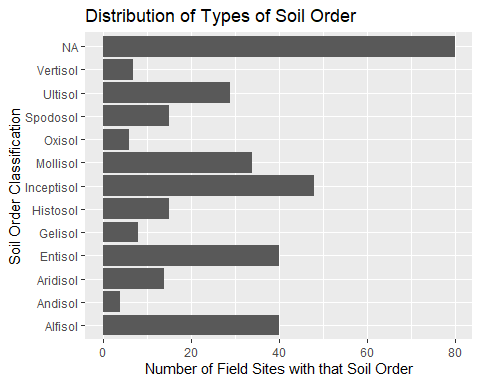
czenSitesNetworks %>% group\_by(Networks) %>% count(Networks)

## # A tibble: 26 x 2  
## # Groups: Networks [26]  
## Networks n  
## <chr> <int>  
## 1 ARS 18  
## 2 CERN 14  
## 3 CRITEX 16  
## 4 CUAHSI 4  
## 5 CZO 14  
## 6 CZO Affiliate 6  
## 7 DOE 6  
## 8 EarthShape 4  
## 9 H+ 5  
## 10 Independent 18  
## # ... with 16 more rows

Soil Order:

Oxisol and Andisols are the least representative soil orders found at the research sites. This is supportive of our earlier conclusisions - According to the USDA, oxisols are found in tropical and subtropical enviroments and are not naturally fertile. Additionally, according to the USDA, andisols are compartively rare and only make up about 1% of the World’s surface. (USDA, <https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/edu/?cid=nrcs142p2_053588>).

ggplot(data = czenSitesSoilOrder) +  
 geom\_bar(mapping = aes(x = SoilOrder)) +  
 coord\_flip() +  
 labs(title = "Distribution of Types of Soil Order", y = "Number of Field Sites with that Soil Order", x = "Soil Order Classification")



czenSitesSoilOrder %>% group\_by(SoilOrder) %>% count(SoilOrder)

## # A tibble: 13 x 2  
## # Groups: SoilOrder [13]  
## SoilOrder n  
## <chr> <int>  
## 1 Alfisol 40  
## 2 Andisol 4  
## 3 Aridisol 14  
## 4 Entisol 40  
## 5 Gelisol 8  
## 6 Histosol 15  
## 7 Inceptisol 48  
## 8 Mollisol 34  
## 9 Oxisol 6  
## 10 Spodosol 15  
## 11 Ultisol 29  
## 12 Vertisol 7  
## 13 <NA> 80

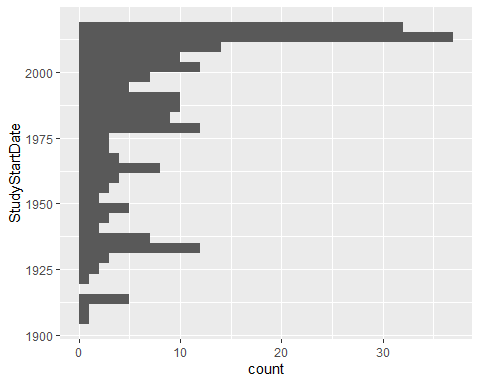
Study Start Date:

Most long-term research has begun since the major environmental legislation was passed by the Nixon administration. With increases in access to information and focus on how anthropogenic climate change is affecting the Earth’s environmental systems, scientists are becoming more focused on tracking the changes of these systems over longer stretches of time.

ggplot(data = czenSites) +  
 geom\_histogram(mapping = aes(x = StudyStartDate)) +  
 coord\_flip()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 6 rows containing non-finite values (stat\_bin).



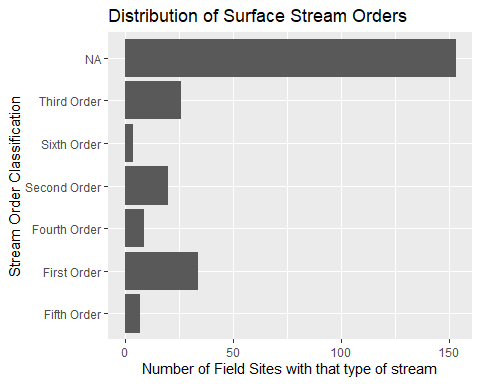
summary(czenSites$StudyStartDate)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1907 1965 1996 1986 2014 2018 6

Stream Order

Stream order is a measurment of how many tributaries flow into a stream or river. First order streams are streams who’s primary source of water is from the landscape and/or a spring. When first order streams combine, they produce second order streams. Therefore, this figure is logical in that it shows an abundance of first, second, and third order streams, and less fourth, fifth, and sixth order streams.

ggplot(data = czenSitesStreamOrder) +  
 geom\_bar(mapping = aes(x = SurfaceStreamOrder)) +  
 coord\_flip() +  
 labs(title = "Distribution of Surface Stream Orders", y = "Number of Field Sites with that type of stream", x = "Stream Order Classification")



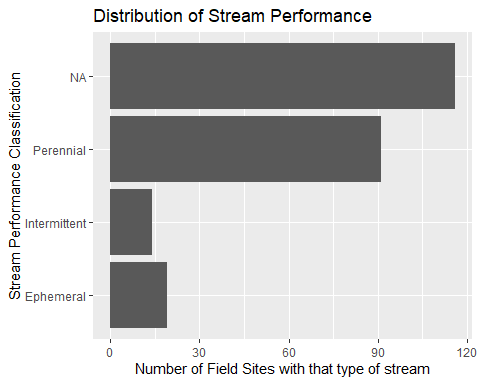
czenSitesStreamOrder %>% group\_by(SurfaceStreamOrder) %>% count(SurfaceStreamOrder)

## # A tibble: 7 x 2  
## # Groups: SurfaceStreamOrder [7]  
## SurfaceStreamOrder n  
## <chr> <int>  
## 1 Fifth Order 7  
## 2 First Order 34  
## 3 Fourth Order 9  
## 4 Second Order 20  
## 5 Sixth Order 4  
## 6 Third Order 26  
## 7 <NA> 153

Stream Flow Performance

Most streams that are studied or present at research sites are perennial streams.

ggplot(data = czenSitesStreamPerformance) +  
 geom\_bar(mapping = aes(x = StreamFlowPerformance)) +  
 coord\_flip() +  
 labs(title = "Distribution of Stream Performance", y = "Number of Field Sites with that type of stream", x = "Stream Performance Classification")



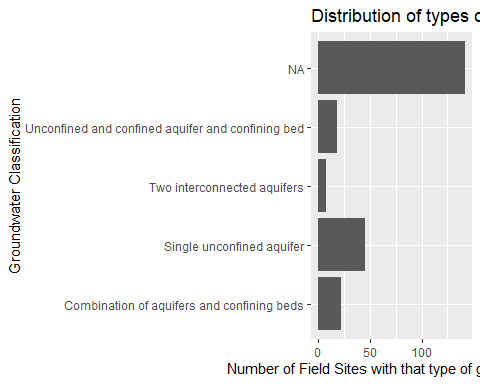
czenSitesStreamPerformance %>% group\_by(StreamFlowPerformance) %>% count(StreamFlowPerformance)

## # A tibble: 4 x 2  
## # Groups: StreamFlowPerformance [4]  
## StreamFlowPerformance n  
## <chr> <int>  
## 1 Ephemeral 19  
## 2 Intermittent 14  
## 3 Perennial 91  
## 4 <NA> 116

Ground water

It appears as though most of the research sites are not focused on measuring groundwater (geologic focus).

ggplot(data = czenSitesGroundwater) +  
 geom\_bar(mapping = aes(x = Groundwater)) +  
 coord\_flip() +  
 labs(title = "Distribution of types of Groundwater", y = "Number of Field Sites with that type of groundwater", x = "Groundwater Classification")



czenSitesGroundwater %>% group\_by(Groundwater) %>% count(Groundwater)

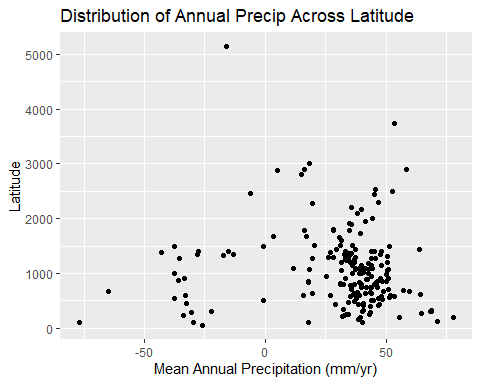
## # A tibble: 5 x 2  
## # Groups: Groundwater [5]  
## Groundwater n  
## <chr> <int>  
## 1 Combination of aquifers and confining beds 22  
## 2 Single unconfined aquifer 45  
## 3 Two interconnected aquifers 7  
## 4 Unconfined and confined aquifer and confining bed 18  
## 5 <NA> 142

Covariation

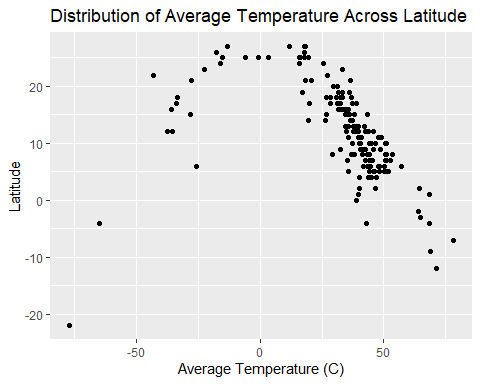
Latitude x Mean Annual Precip & Latitude x AvgTemp (Quantitative x Quantitative)

The most interesting figure to me is the Temperature across Latitude. You can see the decrease in average temperature as latitude diverges from the Equator.

ggplot(data = czenSites) +  
 geom\_point(aes(x = Latitude, y = MeanAnnualPrecip), na.rm = TRUE) +  
 labs(title = "Distribution of Annual Precip Across Latitude", y = "Latitude", x = "Mean Annual Precipitation (mm/yr)")

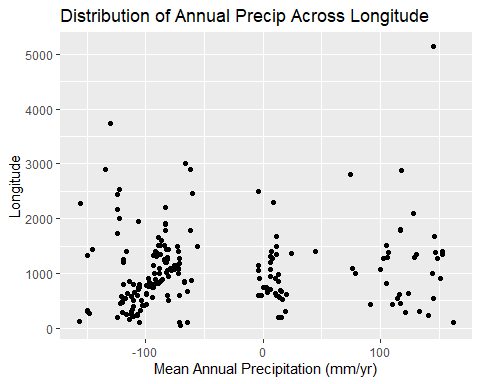


ggplot(data = czenSites) +  
 geom\_point(aes(x = Latitude, y = AvgTemp), na.rm = TRUE) +  
 labs(title = "Distribution of Average Temperature Across Latitude", y = "Latitude", x = "Average Temperature (C)")

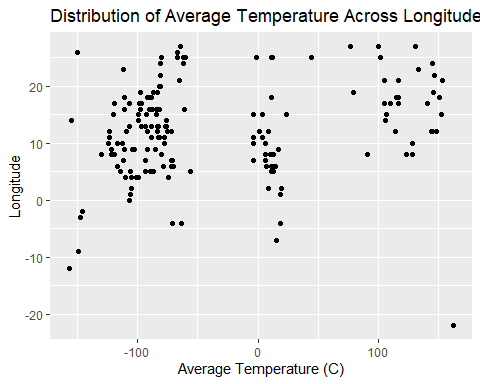


Longitude x Precip & Longitude x Avg Temp (Quantitative x Quantitative)

ggplot(data = czenSites) +  
 geom\_point(aes(x = Longitude, y = MeanAnnualPrecip), na.rm = TRUE) +  
 labs(title = "Distribution of Annual Precip Across Longitude", y = "Longitude", x = "Mean Annual Precipitation (mm/yr)")



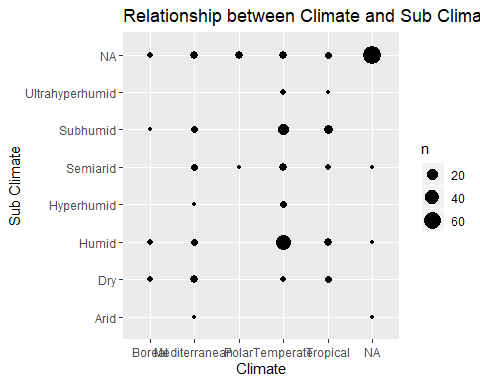
ggplot(data = czenSites) +  
 geom\_point(aes(x = Longitude, y = AvgTemp), na.rm = TRUE)+  
 labs(title = "Distribution of Average Temperature Across Longitude", y = "Longitude", x = "Average Temperature (C)")



Climate x Sub Climate (Categorical x Categorical)

This figure supports the claim that temperate climates are well represented, whereas polar and boreal climates are not.

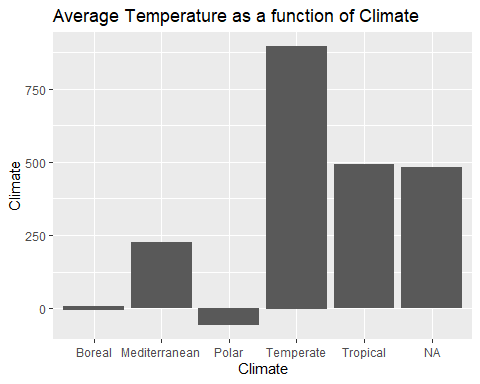
ggplot(data = czenSitesSubClimate) +  
 geom\_count(aes(x = Climate, y = SubClimate)) +  
 labs(title = "Relationship between Climate and Sub Climate", y = "Sub Climate", x = "Climate")



Climate x AvgTemp and Climate x MeanAnnualPrecip (Categorical x Quantitative)

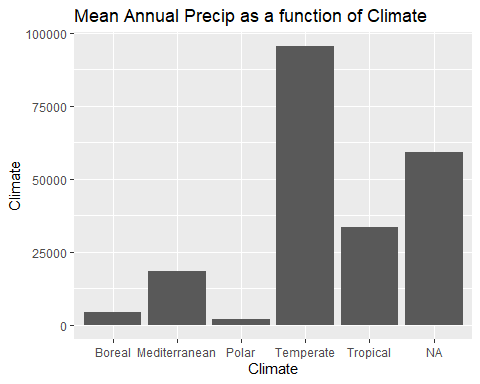
ggplot(data = czenSites) +  
 geom\_col(aes(x = Climate, y = AvgTemp), na.rm = TRUE) +  
 labs(title = "Average Temperature as a function of Climate", y = "Climate", y = "Average Temperature (C)")

## Warning: Removed 63 rows containing missing values (position\_stack).



ggplot(data = czenSites) +  
 geom\_col(aes(x = Climate, y = MeanAnnualPrecip), na.rm = TRUE) +  
 labs(title = "Mean Annual Precip as a function of Climate", y = "Climate", y = "Mean Annual Precip (mm/yr)")

## Warning: Removed 30 rows containing missing values (position\_stack).



Reflections The method of data exploration that I chose to pursue was that of fixing known issues right away (I knew I would need to remove the units from AvgTemp and MeanAnnualPrecip). Then I created rough figures for each variable and identified issues that needed to be resolved. These included misspellings of data values (secimentary instead of sedimentary, or interntaional instead of international). I then needed to figure out what to do with the variables that had multiple values for some observations. Based on some internet searching, the best course of action I found was to use the tidyr function ‘unnest’ to split the multivalues at the commas that separated them. What this did was create a new observation for each value in the column. Therefore, I needed to create separate dataframes for each multiple value variable I fixed, or else the main czenSites dataframe would have redundant observations. This process was definitely challenging, but rewarding once I figured it out. Once I had completed these preliminary data cleanings, I returned to the individual figures to identify more subtle issues. Once I had identified those issues, I made the changes back in the data cleaning section. I am generally happy with the way my figures turned out. I included multiple covariation figures to illustrate noisy data vs the cleaner Latitude x AvgTemp. I wish I had more knowledge to use color to highlight certain aspects of each figure. Additionally, I tried to figure out how to re-order the Stream Order figure (First -> Sixth instead of Alphabetical) but was unable to figure that out. Overall, I was not surprised by the areas of study that were lacking in representation. Most existing Field sites are in the US and cover a broad range of systems, but there are still many unique systems across the Earth that do not have long-term research organizations established.