US Wildfires and Drought

Trends Over The Last 30 Years

Richard Railton 11/18/21

Introduction

- Why Wildfires?
- Aims:
  - What are the Wildfire trends in the US over the last 30 years?
  - Is Wildfire frequency, extent or severity related to drought?
  - Can I predict Wildfire frequency, extent or severity?

Dataset Overview

Wildfires and Drought

- Primary Dataset: US Wildfires
  - Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617] (5th Edition)[1]
  - 2.17 million geo-referenced wildfire records
- Secondary Datasets: Drought Indicator - Standardized Precipitation Evapotranspiration Index (SPEI)
  - Average Drought Conditions Across the Contiguous 48 States According to the SPEI, 1992–2020[2]
  - Average Drought Conditions Across California According to the SPEI, 1992–2020[3]

[1]Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5
[2]Figure 2. Average Drought Conditions Across the Contiguous 48 States According to the SPEI, 1900–2020 https://www.epa.gov/climate-indicators/climate-change-indicators-drought
[3]https://wrcc.dri.edu/wwdt/time/

Wildfires Overview

- 2.17 million geo-referenced wildfire records from 1992-2018
- Downloaded as a SQLITE Database and imported
- Data transformation and cleaning required

glimpse(fires)

```
Rows: 2,166,753
Columns: 37
$ FOD_ID                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28,…
$ FPA_ID                    <chr> "FS-1418826", "FS-1418827", "FS-1418835", "FS-1418845", "FS-1418847", "FS-1418849", "FS-1418851", "FS-…
$ SOURCE_SYSTEM_TYPE        <chr> "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED", "FED…
$ SOURCE_SYSTEM             <chr> "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT", "FS-FIRESTAT…
$ NWCG_REPORTING_AGENCY     <chr> "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", "FS", …
$ NWCG_REPORTING_UNIT_ID    <chr> "USCAPNF", "USCAENF", "USCAENF", "USCAENF", "USCAENF", "USCAENF", "USCAENF", "USCASHF", "USCASHF", "US…
$ NWCG_REPORTING_UNIT_NAME  <chr> "Plumas National Forest", "Eldorado National Forest", "Eldorado National Forest", "Eldorado National F…
$ SOURCE_REPORTING_UNIT     <int> 511, 503, 503, 503, 503, 503, 503, 514, 514, 503, 503, 503, 503, 503, 503, 503, 503, 503, 308, 30…
$ SOURCE_REPORTING_UNIT_NAME <chr> "Plumas National Forest", "Eldorado National Forest", "Eldorado National Forest", "Eldorado National F…
$ LOCAL_FIRE_REPORT_ID      <int> 1, 13, 27, 43, 44, 54, 58, 3, 5, 61, 64, 71, 91, 99, 102, 103, 109, 111, 115, 8, 12, 13, 16, 19, 20, 2…
$ LOCAL_INCIDENT_ID         <chr> "PNF-47", "13", "021", "6", "7", "8", "9", "02", "03", "10", "11", "14", "26", "21", "28", "22", "23",…
$ FIRE_CODE                 <chr> "BJ8K", "AAC0", "A32W", NA, NA, NA, NA, "BK5X", "BLPQ", NA, NA, NA, NA, NA, NA, NA, NA, "BHA3", NA, "A…
$ FIRE_NAME                 <chr> "FOUNTAIN", "PIGEON", "SLACK", "DEER", "STEVENOT", "HIDDEN", "FORK", "SLATE", "SHASTA", "TANGLEFOOT", …
$ ICS_209_PLUS_INCIDENT_JOIN_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "2004_CA-ENF-017646_POWER", "2004_CA-E…
$ ICS_209_PLUS_COMPLEX_JOIN_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ MTBS_ID                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "CA3850212028020041006", "CA3878712031…
$ MTBS_FIRE_NAME            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "POWER", "FREDS", NA, NA, NA, NA, …
$ COMPLEX_NAME              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ FIRE_YEAR                 <int> 2005, 2004, 2004, 2004, 2004, 2004, 2004, 2005, 2005, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, …
$ DISCOVERY_DATE            <chr> "2/2/2005 0:00", "5/12/2004 0:00", "5/31/2004 0:00", "6/28/2004 0:00", "6/28/2004 0:00", "6/28/2004 0:00", "6/30/2004 0:…
$ DISCOVERY_DOY             <int> 33, 133, 152, 180, 180, 182, 183, 67, 74, 183, 184, 184, 247, 272, 277, 277, 280, 287, 325, 156, 171, …
$ DISCOVERY_TIME            <int> 1300, 845, 1921, 1600, 1600, 1800, 1800, 1300, 1200, 1800, 1830, 1730, 1600, 1200, 630, 930, 1415, 161…
$ NWCG_CAUSE_CLASSIFICATION <chr> "Human", "Natural", "Human", "Natural", "Natural", "Natural", "Natural", "Human", "Human", "Natural", …
$ NWCG_GENERAL_CAUSE        <chr> "Power generation/transmission/distribution", "Natural", "Debris and open burning", "Natural", "Natura…
$ NWCG_CAUSE_AGE_CATEGORY   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ CONT_DATE                 <chr> "2/2/2005 0:00", "5/12/2004 0:00", "5/31/2004 0:00", "7/3/2004 0:00", "7/3/2004 0:00", "7/1/2004 0:00"…
$ CONT_DOY                  <int> 33, 133, 152, 185, 185, 183, 184, 67, 74, 184, 185, 185, 247, 272, 277, 277, 295, 291, 326, 156, 171, …
$ CONT_TIME                 <int> 1730, 1530, 2024, 1400, 1200, 1600, 1400, 1600, 1700, 1800, 1900, 2000, 1630, 1800, 1500, 1400, 1000, …
$ FIRE_SIZE                 <dbl> 0.10, 0.25, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.80, 1.00, 0.10, 0.10, 0.10, 0.10, 0.10, 6.00, 0.20, 0.10, 16823.…
$ FIRE_SIZE_CLASS           <chr> "A", "A", "A", "A", "A", "A", "A", "B", "B", "A", "A", "A", "A", "A", "B", "A", "A", "G", "G", "A", "A", "A…
$ LATITUDE                  <dbl> 40.03694, 38.93306, 38.98417, 38.55917, 38.55917, 38.63528, 38.68833, 40.96806, 41.23361, 38.54833, 38…
$ LONGITUDE                 <dbl> -121.00583, -120.40444, -120.73556, -119.91333, -119.93306, -120.10361, -120.15333, -122.43389, -122.2…
$ OWNER_DESCR               <chr> "USFS", "USFS", "STATE OR PRIVATE", "USFS", "USFS", "USFS", "USFS", "STATE OR PRIVATE", "STATE OR PRIV…
$ STATE                     <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", …
$ COUNTY                    <chr> "63", "61", "17", "3", "3", "5", "17", NA, NA, "5", "17", "5", "17", "5", "17", "17", "5", "17", "17",…
$ FIPS_CODE                 <chr> "06063", "06061", "06017", "06003", "06003", "06005", "06017", NA, NA, "06005", "06017", "06005", "060…
$ FIPS_NAME                 <chr> "Plumas County", "Placer County", "El Dorado County", "Alpine County", "Alpine County", "Amador County…
```

fires_na_count <- sapply(fires, function(y) sum(length(which(is.na(y)))))
fires_na_count <- data.frame(fires_na_count)
fires_na_count

	fires_na_count
	<int>
FOD_ID	0
FPA_ID	0
SOURCE_SYSTEM_TYPE	0
SOURCE_SYSTEM	0
NWCG_REPORTING_AGENCY	0
NWCG_REPORTING_UNIT_ID	0
NWCG_REPORTING_UNIT_NAME	0
SOURCE_REPORTING_UNIT	0
SOURCE_REPORTING_UNIT_NAME	0
LOCAL_FIRE_REPORT_ID	1701854
LOCAL_INCIDENT_ID	734948
FIRE_CODE	1797096
FIRE_NAME	939607
ICS_209_PLUS_INCIDENT_JOIN_ID	2135993
ICS_209_PLUS_COMPLEX_JOIN_ID	2165833
MTBS_ID	2153848
MTBS_FIRE_NAME	2153848
COMPLEX_NAME	2161081
FIRE_YEAR	0
DISCOVERY_DATE	0
DISCOVERY_DOY	0
DISCOVERY_TIME	754468
NWCG_CAUSE_CLASSIFICATION	1
NWCG_GENERAL_CAUSE	0
NWCG_CAUSE_AGE_CATEGORY	2093127
CONT_DATE	854553
CONT_DOY	854553
CONT_TIME	933151
FIRE_SIZE	0
FIRE_SIZE_CLASS	0
LATITUDE	0
LONGITUDE	0
OWNER_DESCR	0
STATE	0
COUNTY	657235
FIPS_CODE	657235
FIPS_NAME	657236

Wildfires Key Variables

- Dropped columns deemed irrelevant, converted data types, removed unwanted rows and created two new attributes.
  - COUNTY (Renamed FIPS_NAME) - 621,616 rows with “na” counties which I decided to keep.
  - REGION (New Attribute) - Eastern and Western States to enable east vs west comparison.
  - DAYS_TO_CONT (New Attribute) - Number of days from discovery to containment to understand how long fires are burning.
    - Replaced DAYS_TO_CONT > 150 with NA
    - Replaced NA’s with median value grouped on FIRE_SIZE_CLASS and STATE

- Decided to keep rows with NA in following columns and clean later if required:
    - DISCOVERY_TIME
    - CONT_DATE
    - CONT_DOY
    - CONT_TIME
    - COUNTY

glimpse(fires_new)

```
Rows: 2,120,440
Columns: 18
$ FOD_ID                   <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14", "15", "16", "17", "18", "1
$ FIRE_YEAR                <fct> 2005, 2004, 2004, 2004, 2004, 2004, 2004, 2005, 2005, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
$ DISCOVERY_DATE           <date> 2005-02-02, 2004-05-12, 2004-05-31, 2004-06-28, 2004-06-28, 2004-06-30, 2004-07-01, 2005-03-08, 2005
$ DISCOVERY_DOY            <fct> 33, 133, 152, 180, 180, 182, 183, 67, 74, 183, 184, 184, 247, 272, 277, 277, 280, 287, 325, 156, 171,
$ DISCOVERY_TIME           <times> 13:00:00, 08:45:00, 19:21:00, 16:00:00, 16:00:00, 18:00:00, 18:00:00, 13:00:00, 12:00:00, 18:00:00,
$ NWCG_CAUSE_CLASSIFICATION <fct> Human, Natural, Human, Natural, Natural, Natural, Natural, Human, Human, Natural, Natural, Natural, H
$ NWCG_GENERAL_CAUSE       <fct> Power generation/transmission/distribution, Natural, Debris and open burning, Natural, Natural, Natur
$ CONT_DATE                <date> 2005-02-02, 2004-05-12, 2004-05-31, 2004-07-03, 2004-07-03, 2004-07-01, 2004-07-02, 2005-03-08, 2005
$ CONT_DOY                 <fct> 33, 133, 152, 185, 185, 183, 184, 67, 74, 184, 185, 185, 247, 272, 277, 277, 295, 291, 326, 156, 171,
$ CONT_TIME                <times> 17:30:00, 15:30:00, 20:24:00, 14:00:00, 12:00:00, 16:00:00, 14:00:00, 16:00:00, 17:00:00, 18:00:00,
$ FIRE_SIZE                <dbl> 0.10, 0.25, 0.10, 0.10, 0.10, 0.10, 0.10, 0.80, 1.00, 0.10, 0.10, 0.10, 0.10, 6.00, 0.20, 0.10, 16823
$ FIRE_SIZE_CLASS          <fct> A, A, A, A, A, A, A, A, B, B, B, A, A, A, A, B, A, A, G, G, A, A, A, A, A, A, A, B, C, B, A, A, A, A, A, A, A,
$ LATITUDE                 <dbl> 40, 39, 39, 39, 39, 39, 39, 41, 41, 39, 39, 39, 39, 38, 39, 39, 39, 39, 39, 33, 33, 33, 34, 33, 34, 3
$ LONGITUDE                <dbl> -121, -120, -121, -120, -120, -120, -120, -122, -122, -120, -120, -120, -120, -121, -120, -121, -120,
$ STATE                    <fct> CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, CA, NM, NM, NM, NM, NM, NM, N
$ COUNTY                   <fct> Plumas County, Placer County, El Dorado County, Alpine County, Alpine County, Amador County, El Dorad
$ REGION                   <fct> West, West, West, West, West, West, West, West, West, West, West, West, West, West, West, West, West,
$ DAYS_TO_CONT             <dbl> 0, 0, 0, 5, 5, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 15, 4, 1, 0, 0, 1, 1, 1, 5, 0, 0, 0, 0, 0, 0, 1, 1, 0
```

```
fires_new_na_count <- sapply(fires_new, function(y) sum(length(which(is.na(y)))))
fires_new_na_count <- data.frame(fires_new_na_count)
fires_new_na_count
```

	fires_new_na_count <int>
FOD_ID	0
FIRE_YEAR	0
DISCOVERY_DATE	0
DISCOVERY_DOY	0
DISCOVERY_TIME	714898
NWCG_CAUSE_CLASSIFICATION	0
NWCG_GENERAL_CAUSE	0
CONT_DATE	818599
CONT_DOY	818560
CONT_TIME	893634
FIRE_SIZE	0
FIRE_SIZE_CLASS	0
LATITUDE	0
LONGITUDE	0
STATE	0
COUNTY	621616
REGION	0
DAYS_TO_CONT	0

Drought Indicator Variable (SPEI)

- Additional Variables: Average Annual five-year SPEI Index for Contiguous US States and California
- Downloaded as a .csv and imported and merged
- The average SPEI value for each year is based on conditions over the preceding five years (five-year SPEI).
- Positive values represent wetter-than-average conditions, while negative values represent drier-than-average conditions.
    - The SPEI is designed to take into account both precipitation and potential evapotranspiration in determining drought.

glimpse(combined_SPEI)

```
Rows: 30
Columns: 3
$ Year        <fct> 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, …
$ US_5yr_SPEI <dbl> -0.29736869, 0.32903045, 0.44245068, 0.74043491, 0.71709022, 1.06105053, 0.93202471, 0.90186513, 0.41446862, 0.21723003,…
$ CA_5yr_SPEI <dbl> -1.19, -0.52, -0.81, 0.28, 0.36, 0.60, 0.96, 1.25, 0.67, 0.33, -0.05, -0.98, -1.04, -0.73, -0.16, -0.26, -0.53, -0.74, -…
```

glimpse(combined_SPEI_group)

```
Rows: 60
Columns: 3
$ Year     <fct> 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, …
$ SPEI     <dbl> -0.297, 0.329, 0.442, 0.740, 0.717, 1.061, 0.932, 0.902, 0.414, 0.217, -0.387, -0.608, -0.712, -0.256, -0.323, -0.115, -0.181,…
$ Location <chr> "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5yr_SPEI", "US_5y…
```

Dataset Summary

summary(fires_new)

```
    FOD_ID            FIRE_YEAR       DISCOVERY_DATE      DISCOVERY_DOY     DISCOVERY_TIME                                NWCG_CAUSE_CLASSIFICATION                                NWCG_GENERAL_CAUSE
 Length:2120440     2006   : 114321   Min.   :1992-01-01   185    :  15613   Min.   :00:00:00   Human                                    :1670177   Missing data/not specified/undetermined:513433
 Class :character   2000   :  96020   1st Qu.:1999-08-04   186    :  13499   1st Qu.:12:34:00   Missing data/not specified/undetermined: 142132   Debris and open burning                 :504672
 Mode  :character   2011   :  95302   Median :2006-03-04   100    :  10503   Median :14:55:00   Natural                                  : 308131   Arson/incendiarism                      :309760
                    2007   :  93666   Mean   :2005-10-05   101    :  10470   Mean   :14:36:34                                                       Natural                                 :308131
                    1999   :  88851   3rd Qu.:2011-12-18   108    :  10440   3rd Qu.:17:10:00                                                       Equipment and vehicle use               :175330
                    2008   :  87150   Max.   :2018-12-31   83     :  10208   Max.   :23:59:00                                                       Recreation and ceremony                 : 90826
                    (Other):1545130                        (Other):2049707   NA's   :714898                                                         (Other)                                 :218288
   CONT_DATE             CONT_DOY         CONT_TIME          FIRE_SIZE       FIRE_SIZE_CLASS    LATITUDE    LONGITUDE         STATE                        COUNTY          REGION          DAYS_TO_CONT
 Min.   :1992-01-01   185    :   9587   Min.   :00:00:00   Min.   :     0   A: 794755       Min.   :25   Min.   :-125   CA     : 235229   Riverside County :  14989   East:1130159   Min.   :  0
 1st Qu.:2000-06-08   186    :   9061   1st Qu.:13:05:00   1st Qu.:     0   B:1023919       1st Qu.:33   1st Qu.:-110   GA     : 180175   Maricopa County  :  12984   West: 990281   1st Qu.:  0
 Median :2007-04-20   184    :   6726   Median :15:55:00   Median :     1   C: 242174       Median :36   Median : -93   TX     : 167061   Lincoln County   :  11852                  Median :  0
 Mean   :2006-09-09   187    :   6594   Mean   :15:23:35   Mean   :    62   D:  31636       Mean   :37   Mean   : -96   NC     : 123793   Washington County:  11815                  Mean   :  1
 3rd Qu.:2013-07-16   108    :   6478   3rd Qu.:18:10:00   3rd Qu.:     3   E:  15653       3rd Qu.:41   3rd Qu.: -83   FL     :  99356   Jackson County   :  11458                  3rd Qu.:  0
 Max.   :2021-07-24   (Other):1263434   Max.   :23:59:00   Max.   :662700   F:   8561       Max.   :49   Max.   : -67   AZ     :  93417   (Other)          :1435726                  Max.   :150
 NA's   :818599       NA's   : 818560   NA's   :893634                      G:   3742                                      (Other):1221409   NA's             : 621616
```

summary(combined_SPEI)

```
     Year      US_5yr_SPEI       CA_5yr_SPEI
1992   : 1   Min.   :-0.71   Min.   :-1.93
1993   : 1   1st Qu.:-0.18   1st Qu.:-1.02
1994   : 1   Median : 0.02   Median :-0.61
1995   : 1   Mean   : 0.19   Mean   :-0.50
1996   : 1   3rd Qu.: 0.72   3rd Qu.:-0.08
1997   : 1   Max.   : 1.06   Max.   : 1.25
(Other):24   NA's   :1
```

Descriptive Statistics

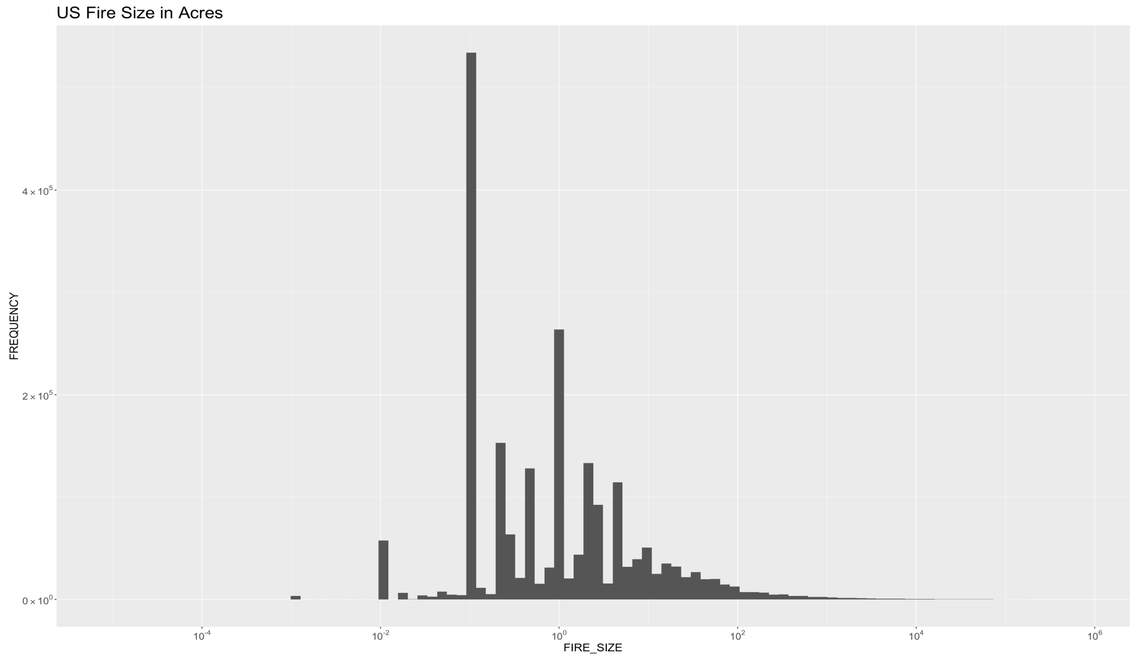
```
attach(fires_new)
fires_new_num <- cbind(FIRE_SIZE, LATITUDE, LONGITUDE, DAYS_TO_CONT)
options(scipen=100)
options(digits=2)
stat.desc(fires_new_num)
```

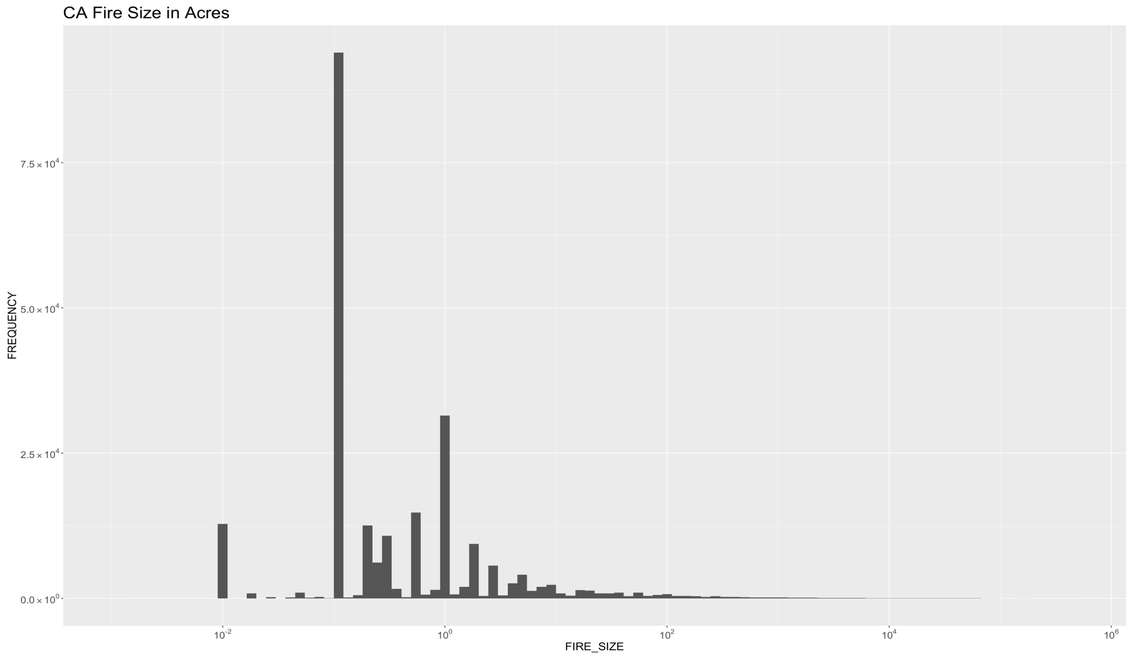
	FIRE_SIZE <dbl>	LATITUDE <dbl>	LONGITUDE <dbl>	DAYS_TO_CONT <dbl>
nbr.val	2120440.00000	2120440.0000	2120440.000	2120440.0000
nbr.null	0.00000	0.0000	0.000	1870956.0000
nbr.na	0.00000	0.0000	0.000	0.0000
min	0.00001	24.5817	-124.719	0.0000
max	662700.00000	49.3434	-66.971	150.0000
range	662699.99999	24.7617	57.748	150.0000
sum	130409712.66672	78415454.4931	-203270757.872	1069271.0000
median	0.97000	35.6909	-93.088	0.0000
mean	61.50125	36.9807	-95.863	0.5043
SE.mean	1.40406	0.0036	0.011	0.0028
CI.mean.0.95	2.75192	0.0071	0.021	0.0054
var	4180231.50750	27.6174	237.263	16.3207
std.dev	2044.56145	5.2552	15.403	4.0399
coef.var	33.24423	0.1421	-0.161	8.0114

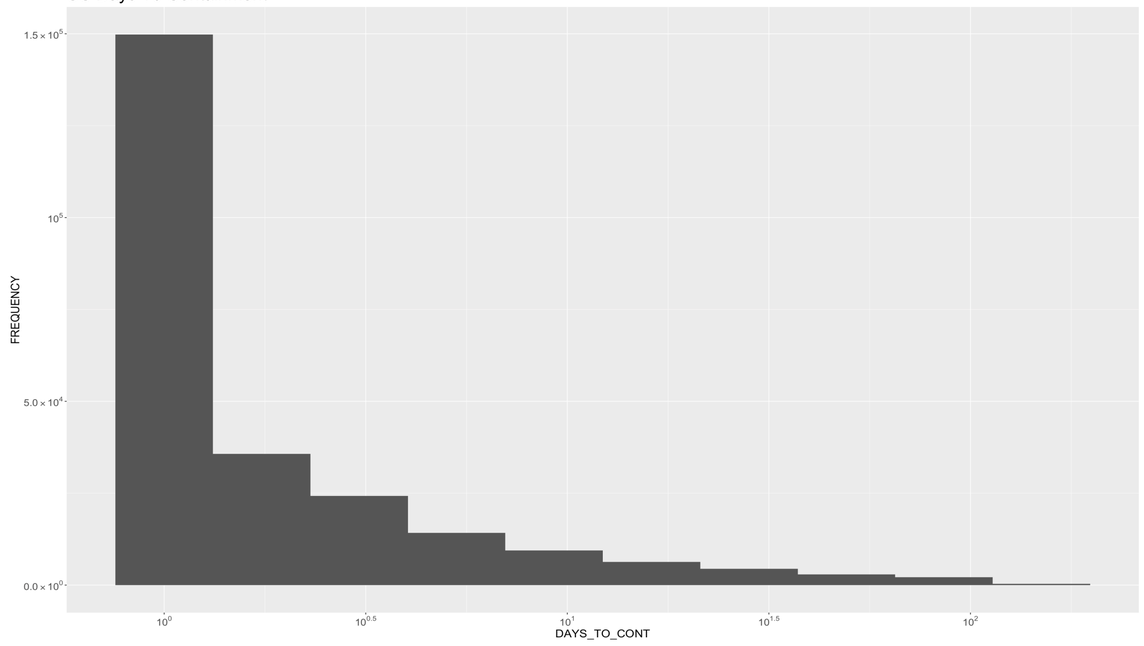
```
attach(combined_SPEI)
combined_SPEI_num <- cbind(US_5yr_SPEI, CA_5yr_SPEI)
stat.desc(combined_SPEI_num)
```

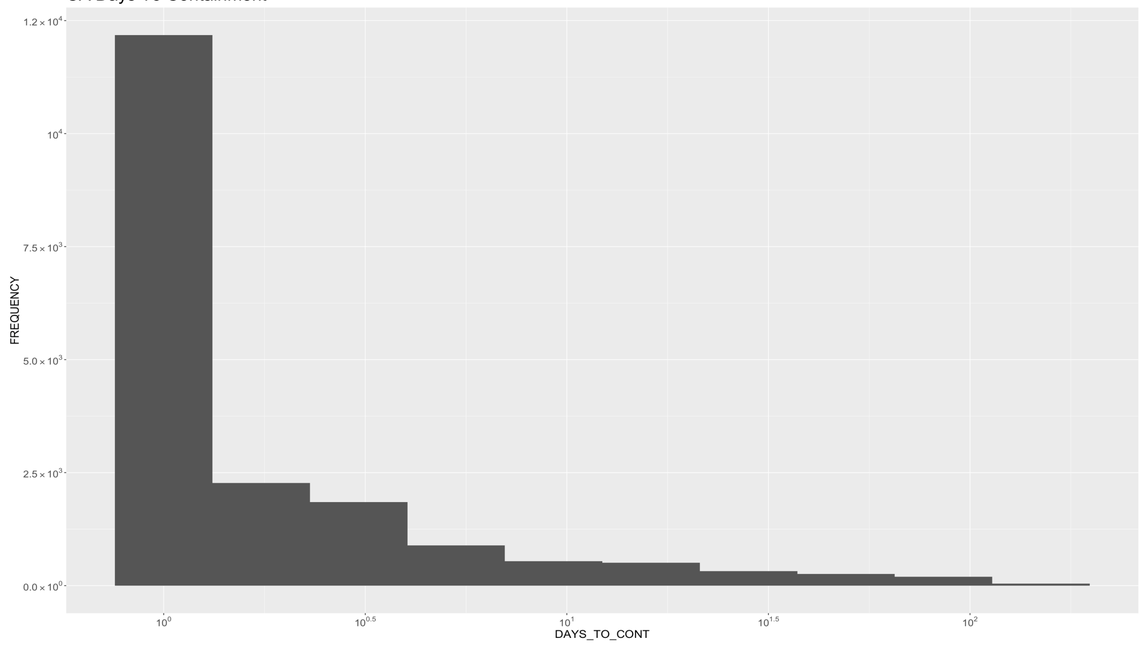
	US_5yr_SPEI <dbl>	CA_5yr_SPEI <dbl>
nbr.val	29.000	30.00
nbr.null	0.000	0.00
nbr.na	1.000	0.00
min	-0.712	-1.93
max	1.061	1.25
range	1.773	3.18
sum	5.614	-14.89
median	0.022	-0.61
mean	0.194	-0.50
SE.mean	0.097	0.14
CI.mean.0.95	0.199	0.29
var	0.273	0.59
std.dev	0.522	0.77
coef.var	2.697	-1.55

Histograms - Fire Size and Days to Containment US & CA

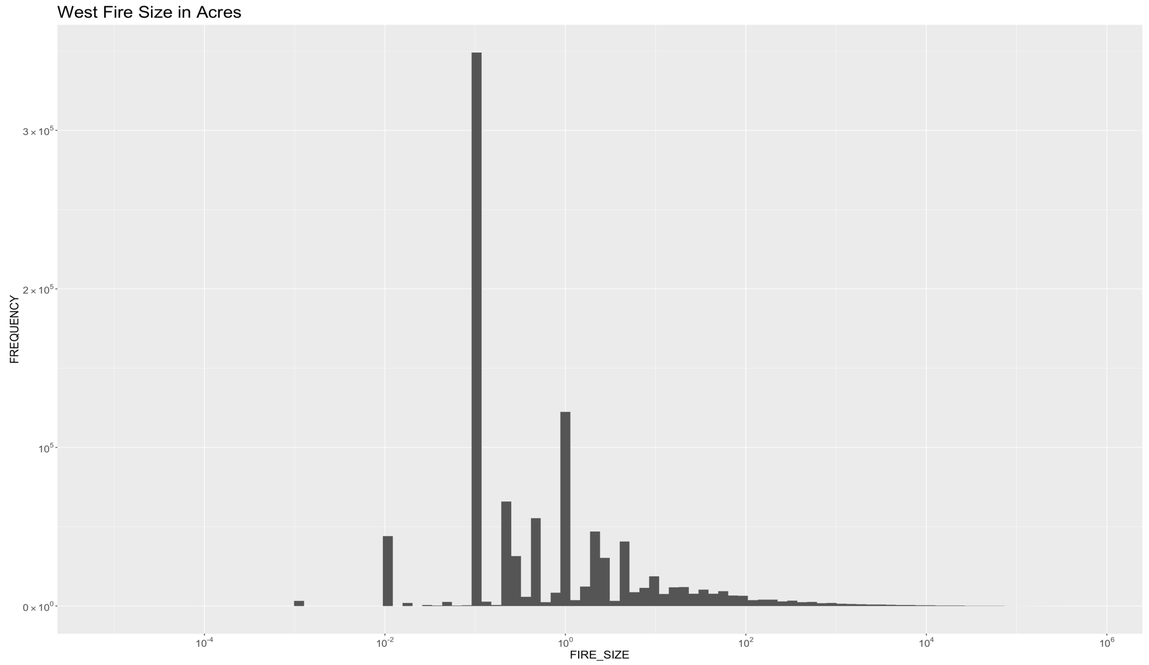


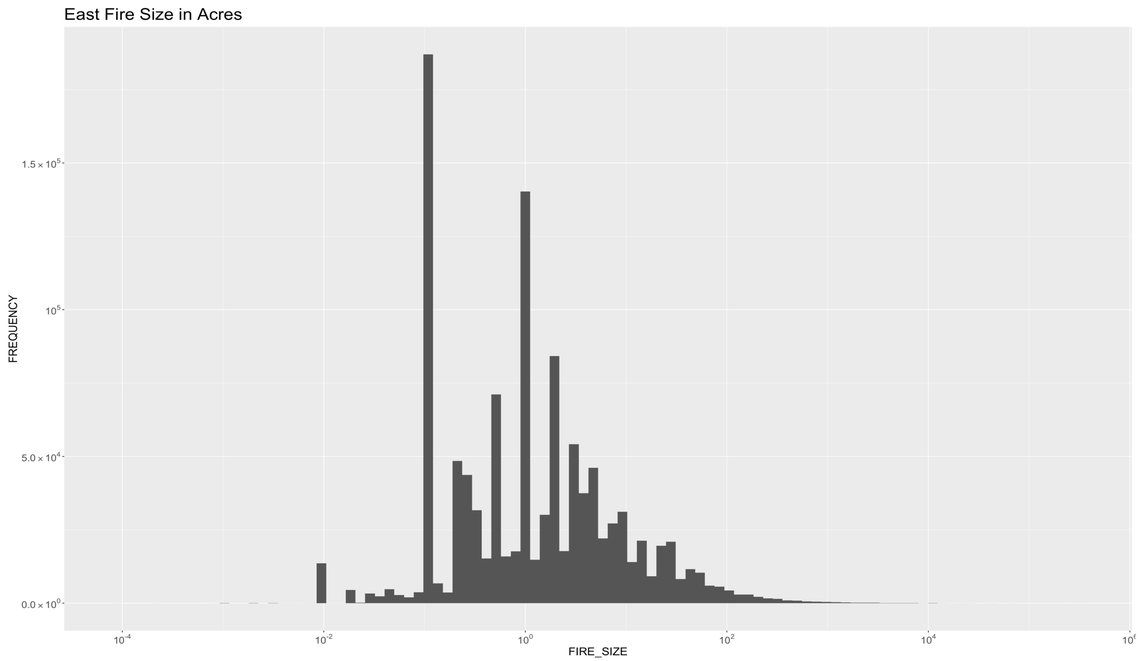


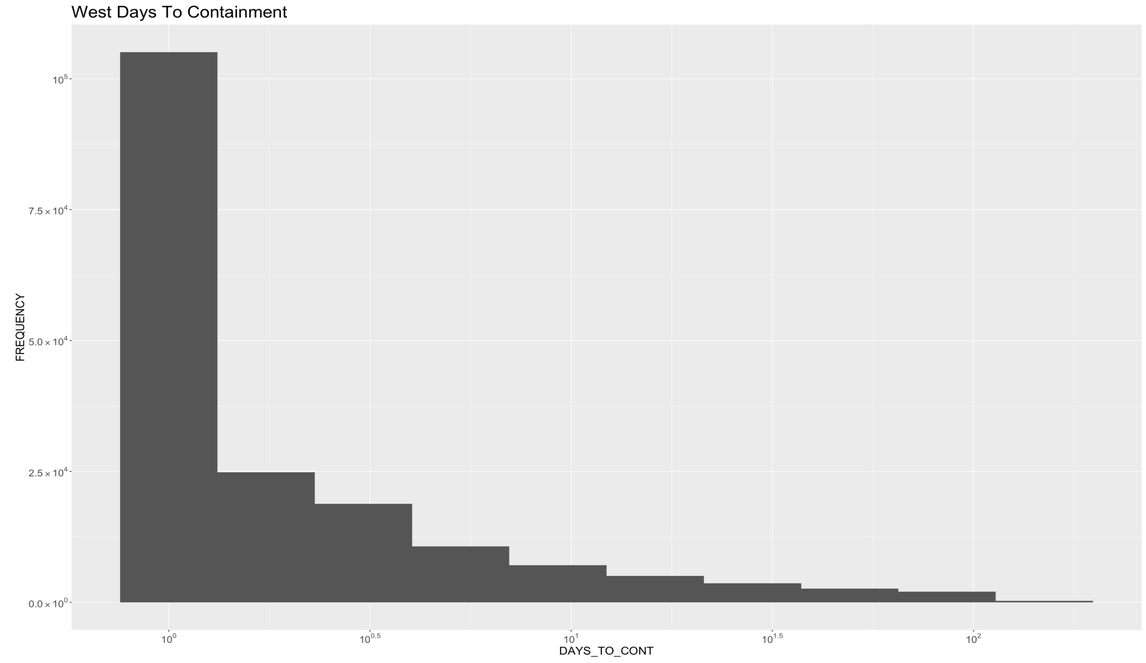


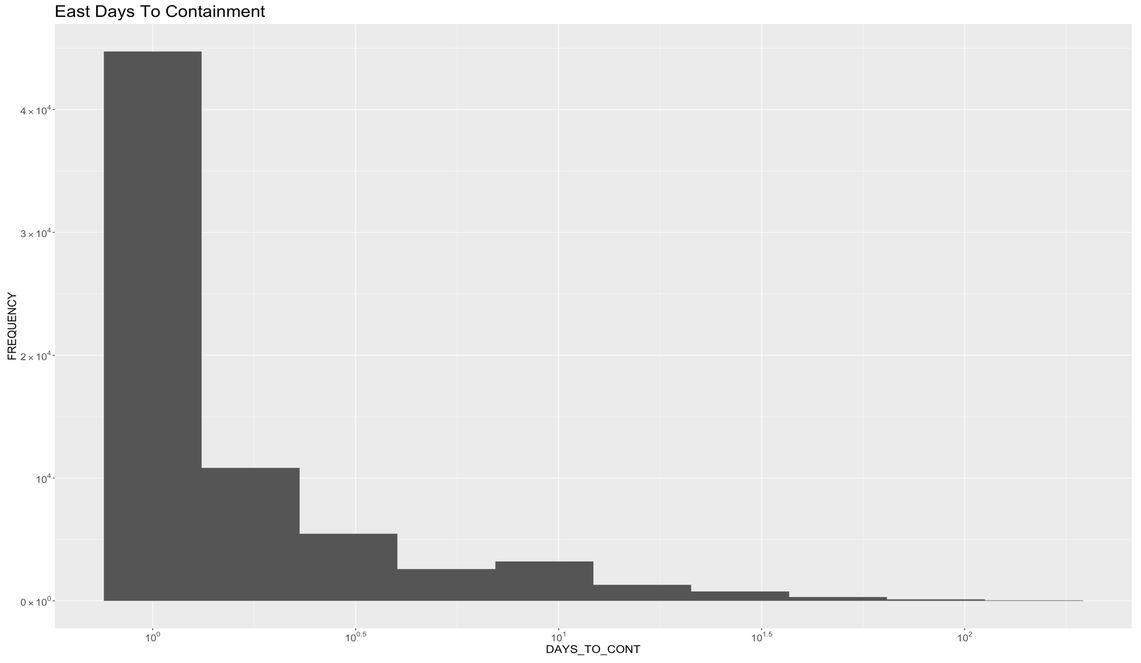


Histograms - Fire Size and Days to Containment West & East US

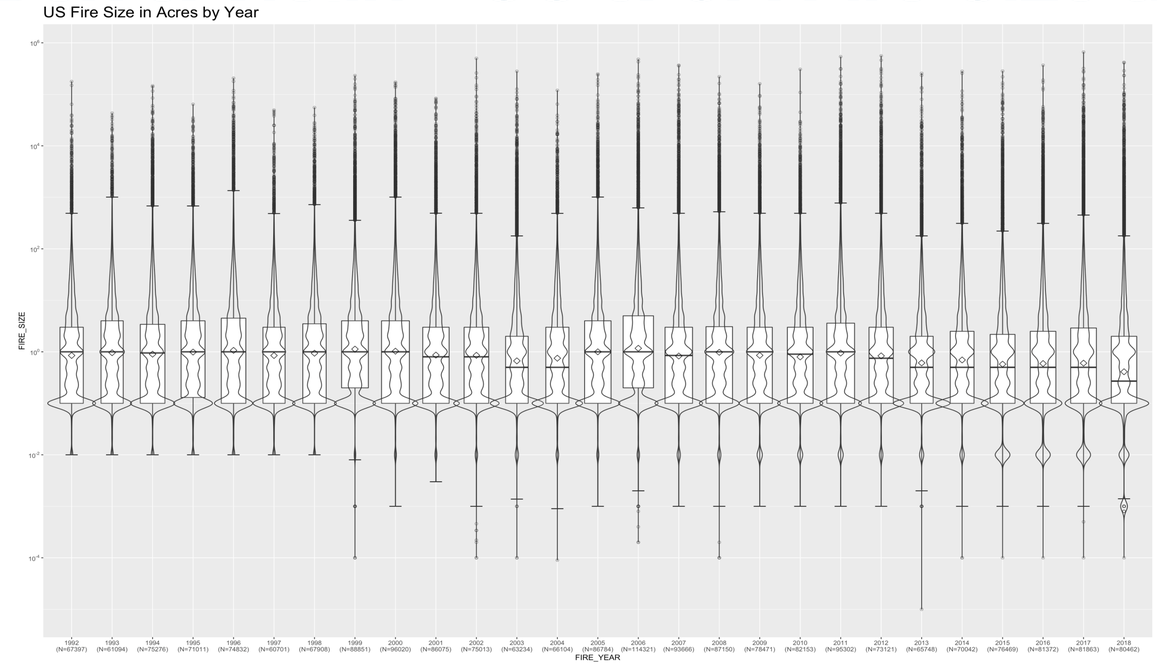


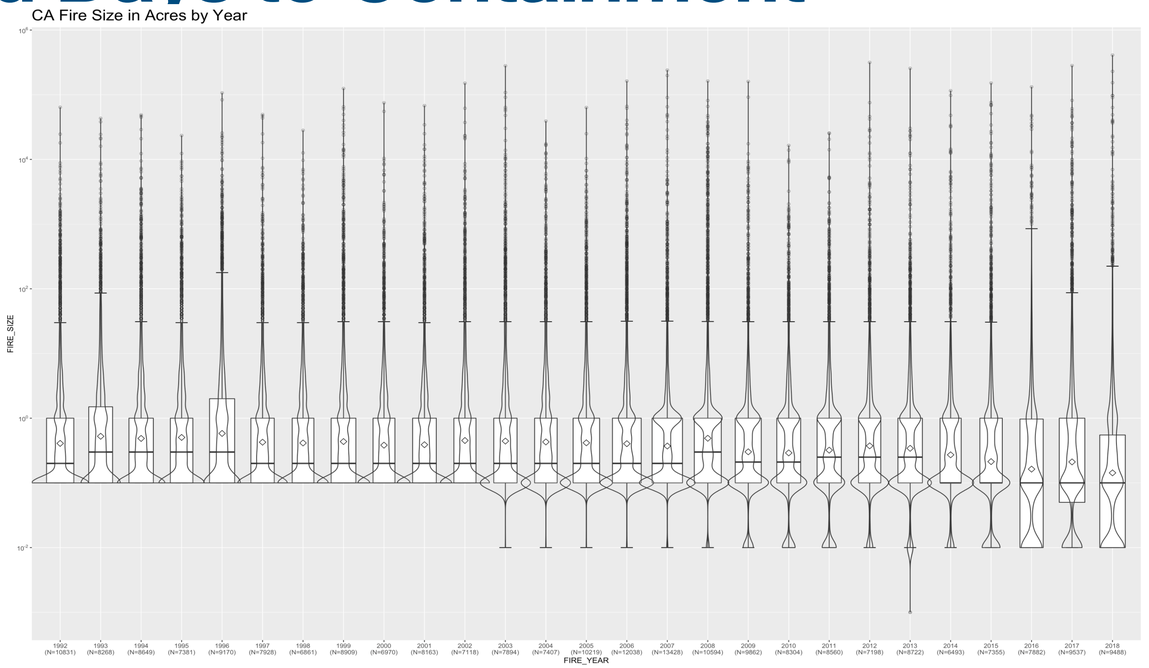


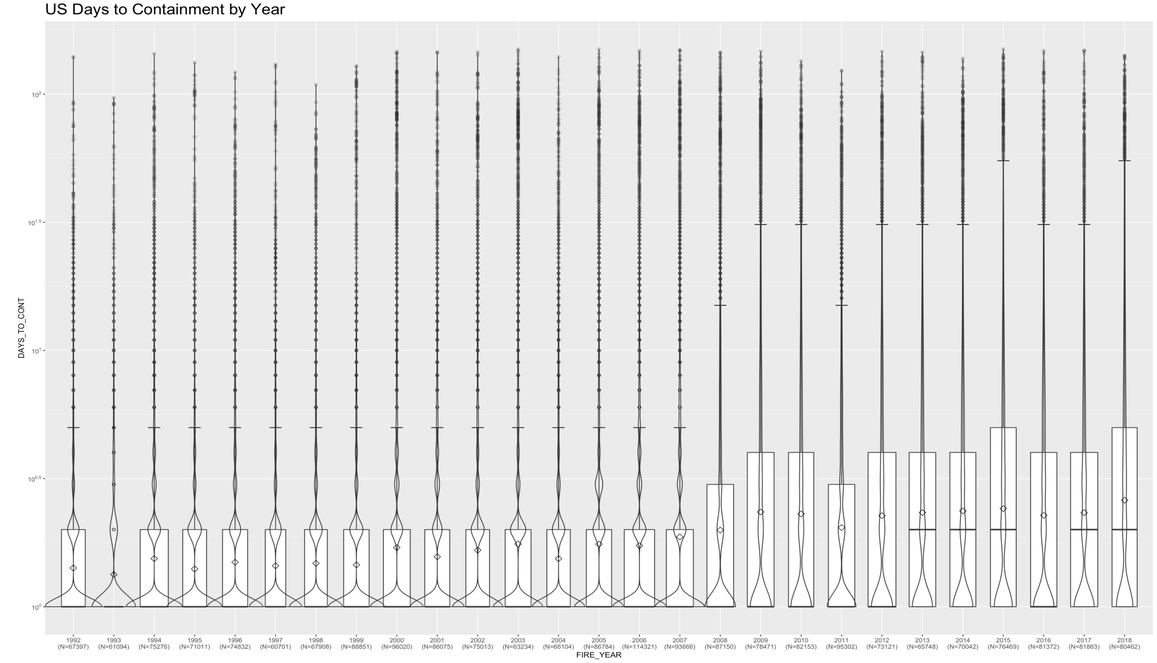


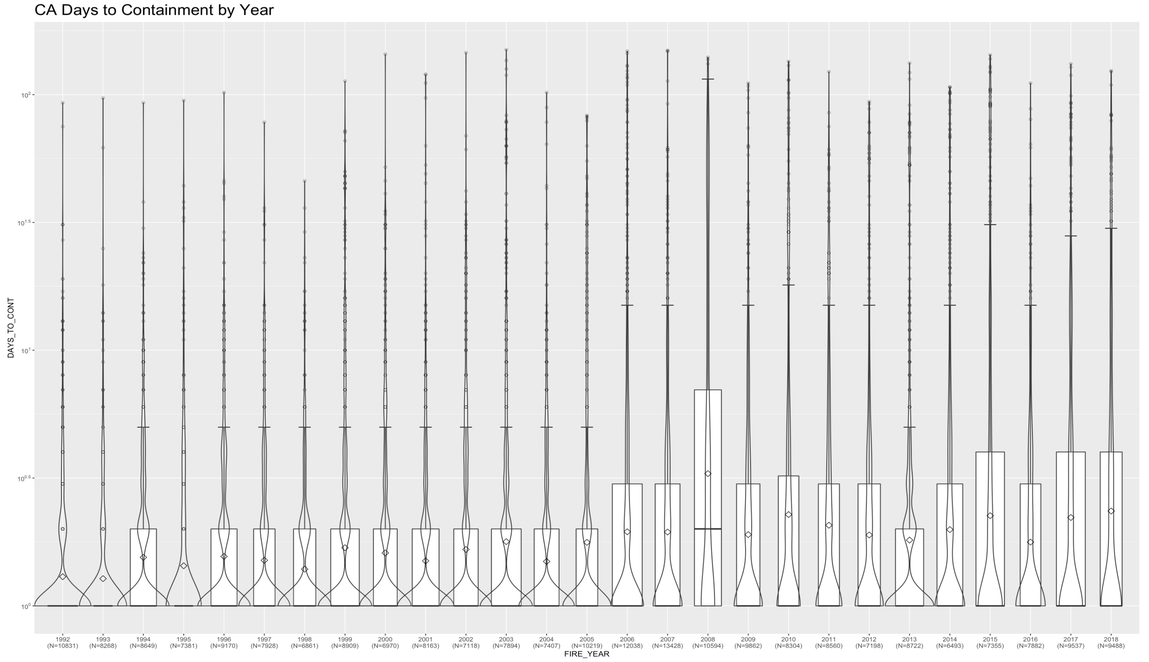


Box & Violin - US & CA Fire Size and Days to Containment

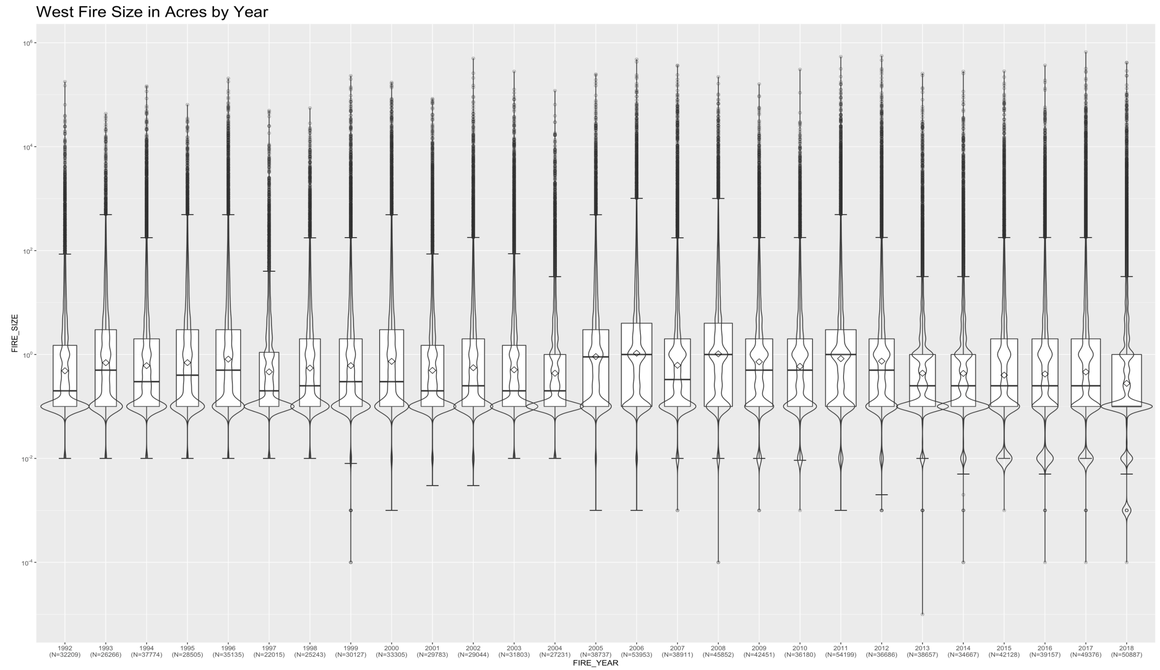


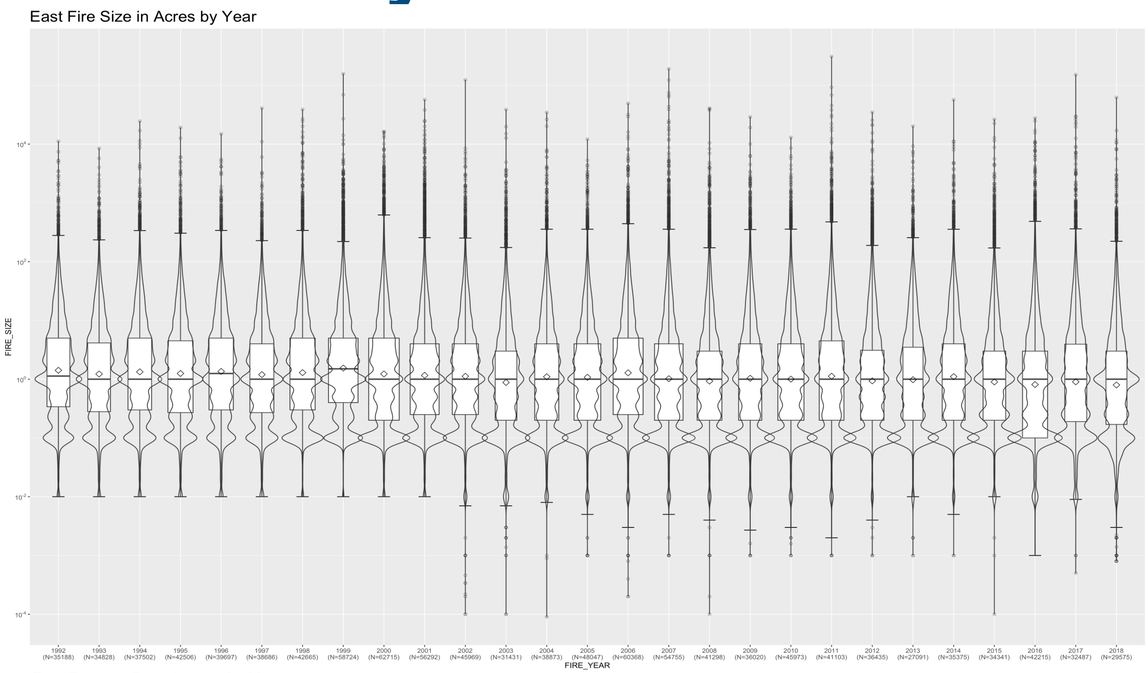


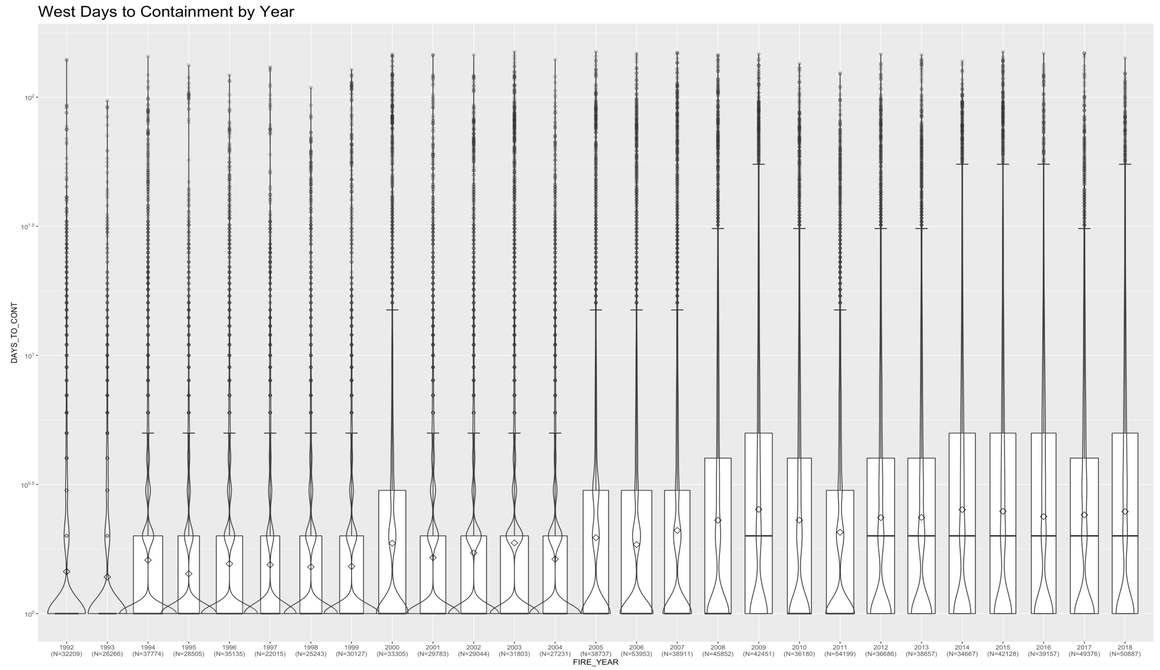


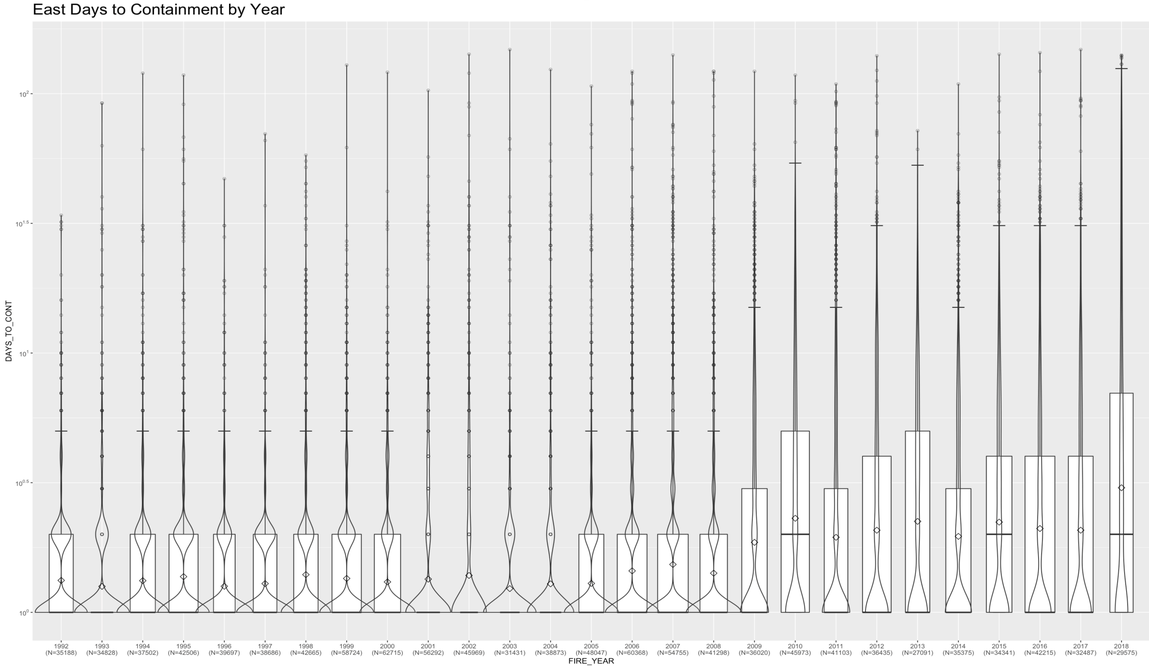


Box & Violin - West & East US Fire Size and Days to Containment

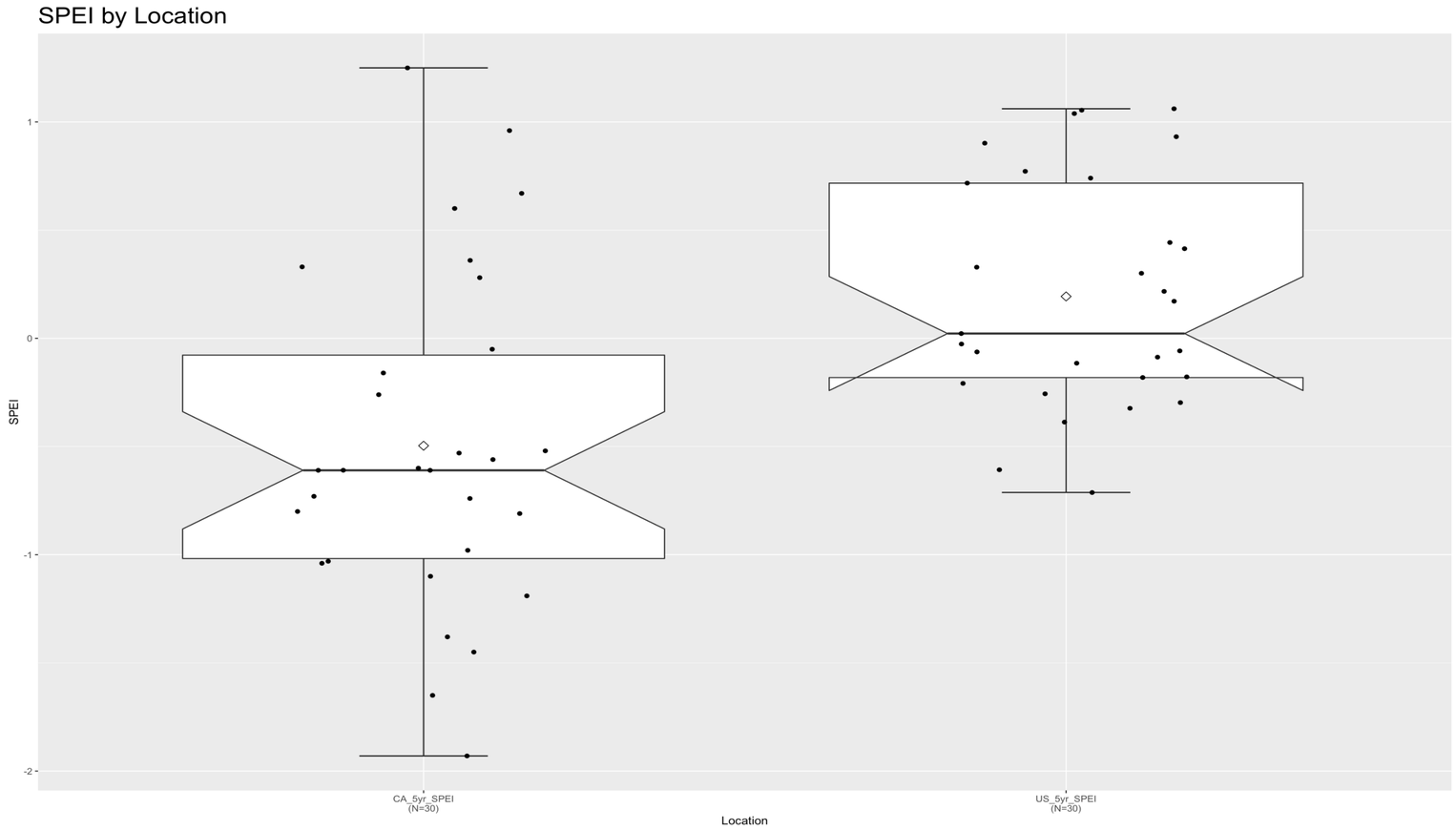






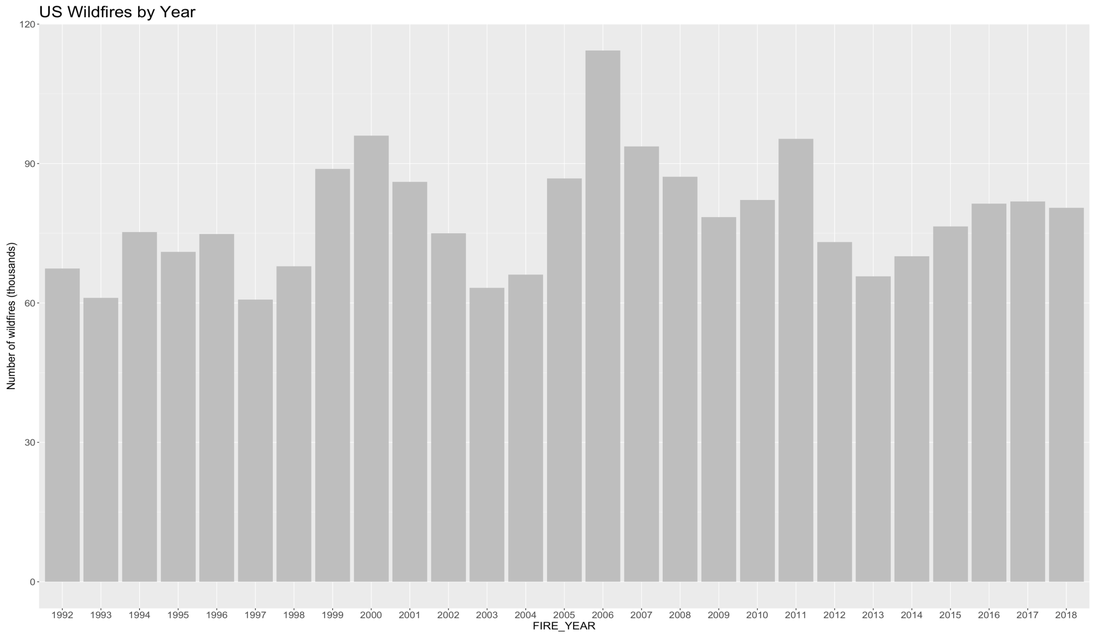


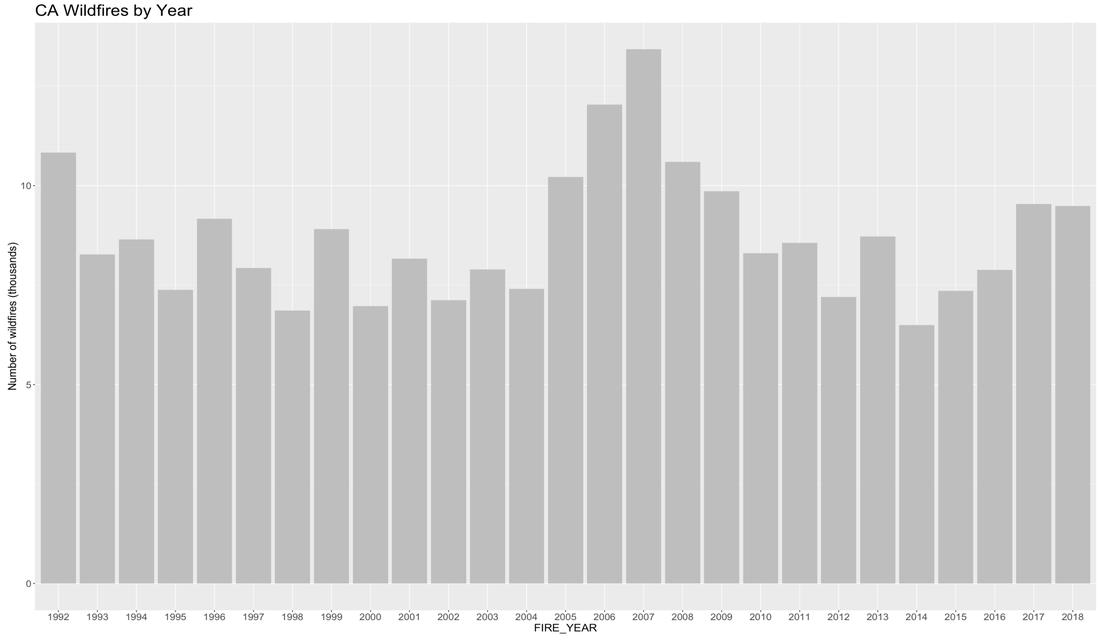
Notched Box Plot - CA & US SPEI

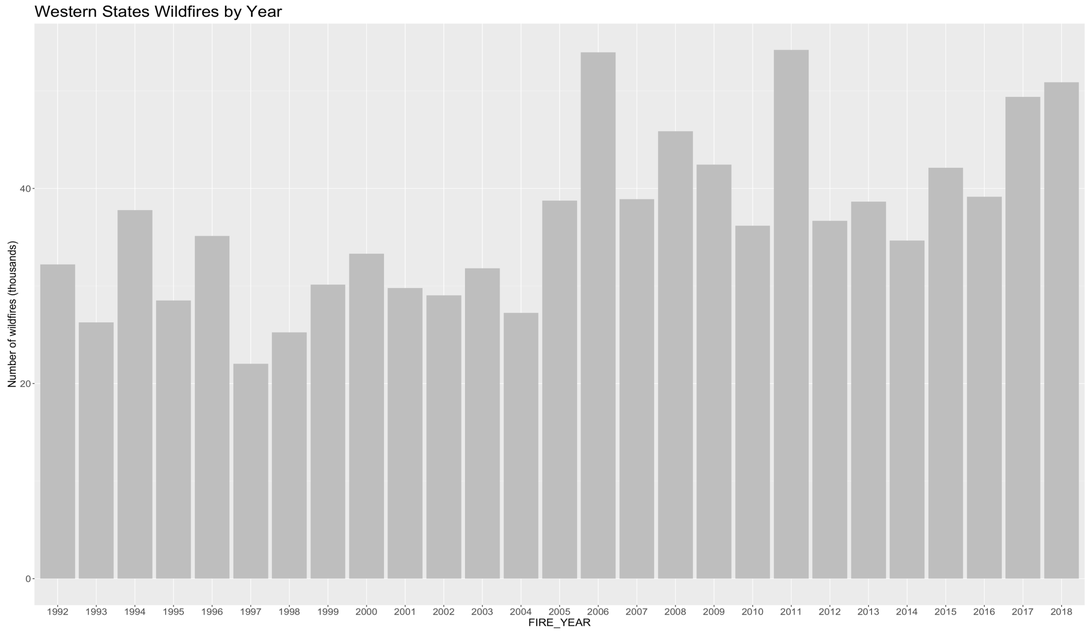


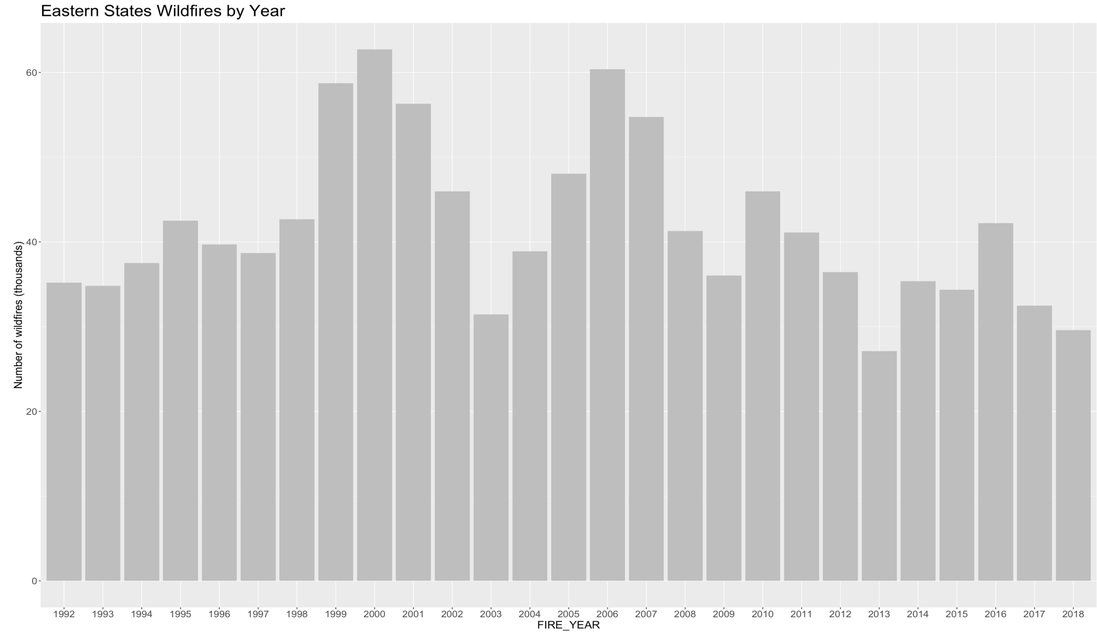
- Since the notches of the two boxes do not overlap, this offers evidence of a statistically significant difference between the medians.
- The CA median 5yr SPEI is negative and also has a greater range indicating California has been getting drier on average over the last 30 years and has more extreme variations in drought conditions.
- The US median 5yr SPEI is almost zero indicating that drought conditions in the US have on average remained in stasis over the last 30 years.
- In order for CA to be drier and the US to stay the same on average over the last 30 years, we can hypothesize that other states may be getting wetter on average.

Fires Over Time US & CA

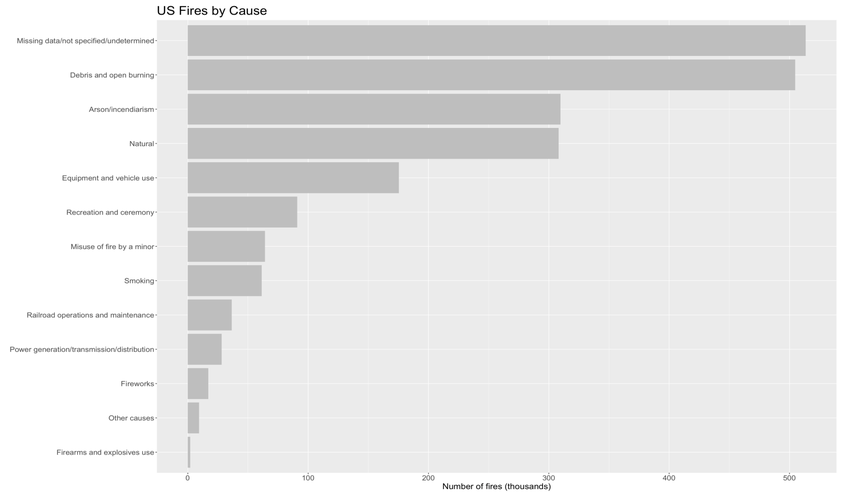


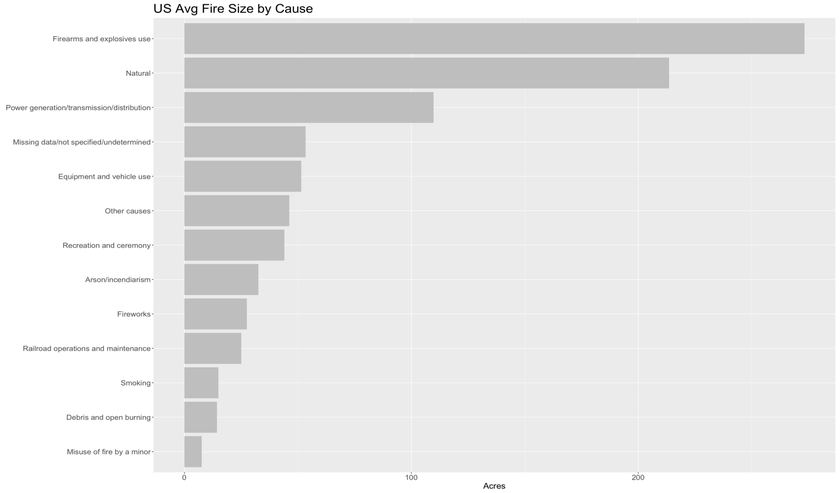


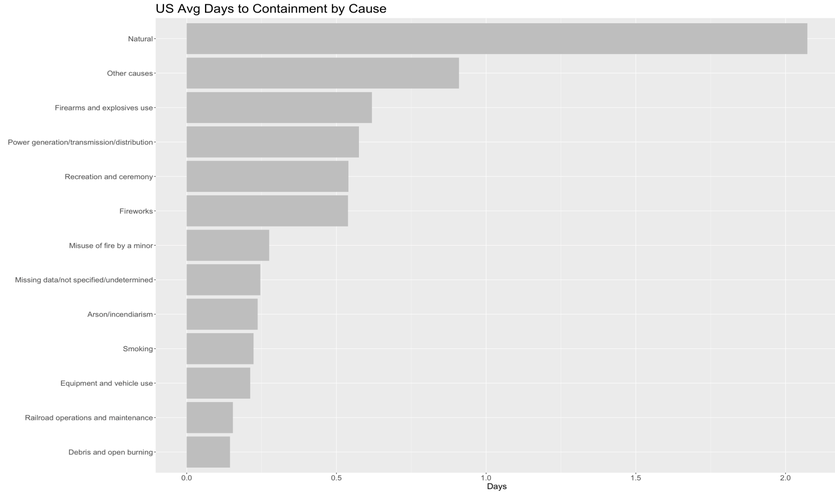


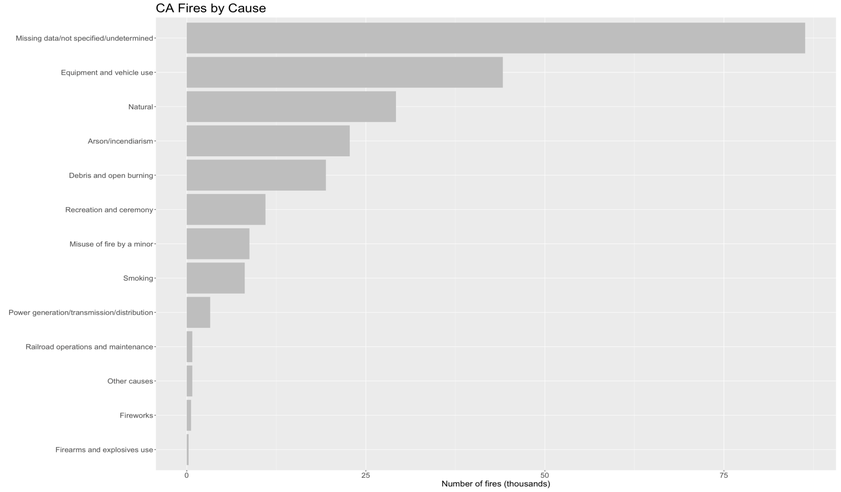


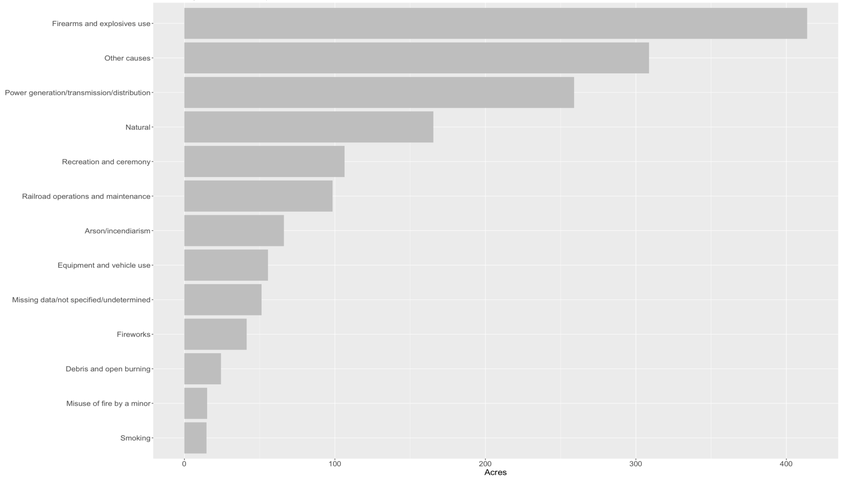
#Fires, Size and Days to Containment by Cause US & CA

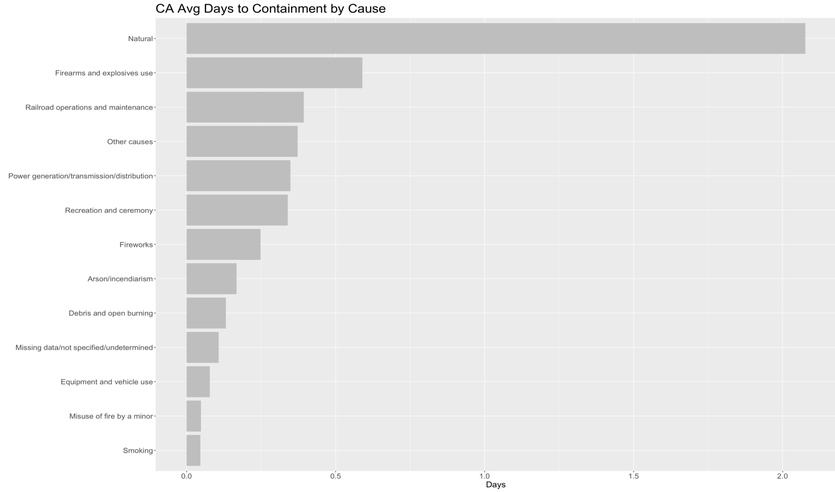




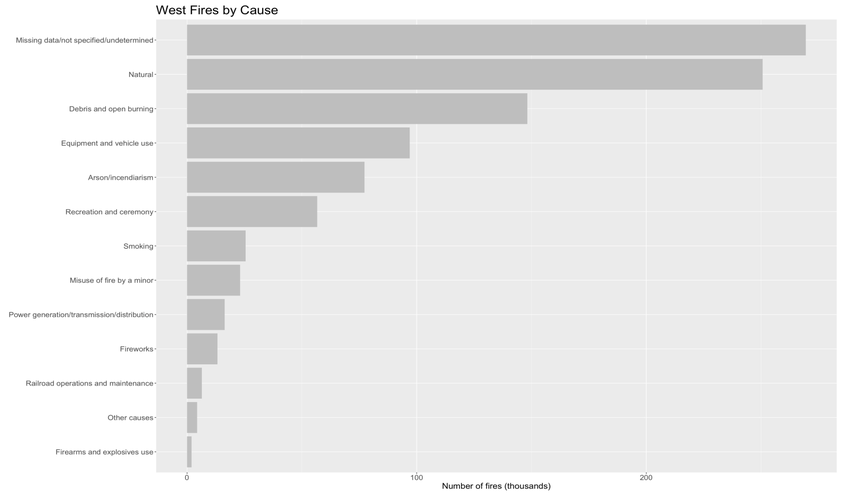


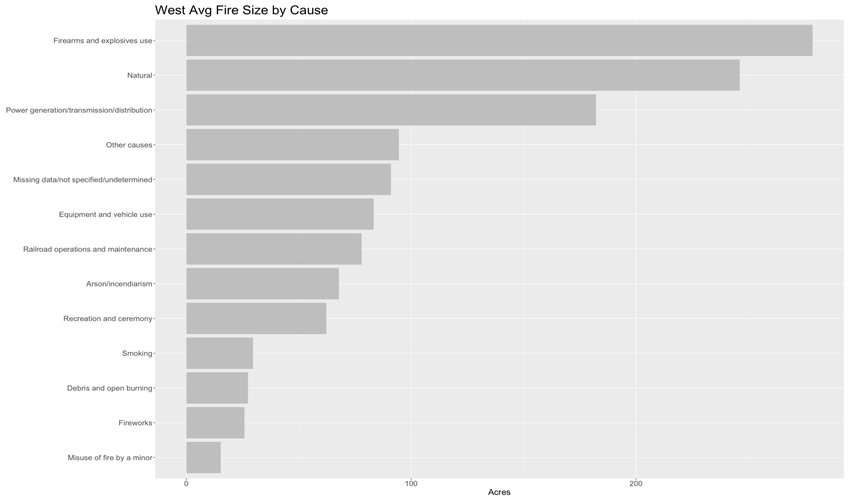


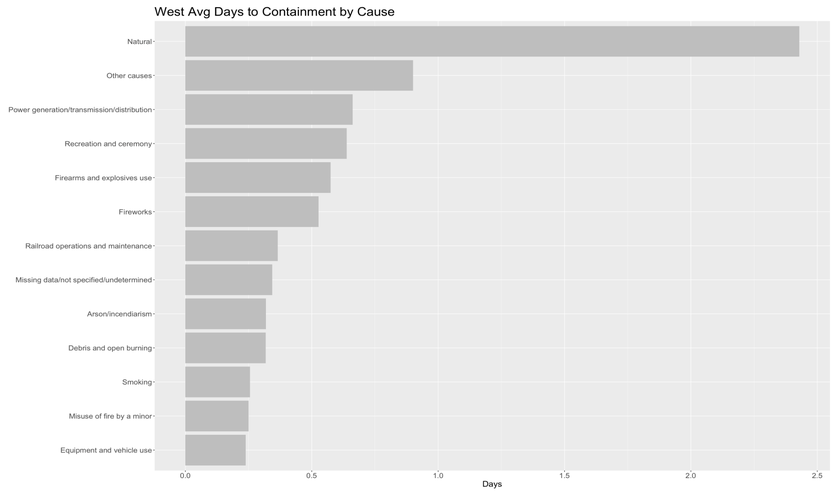


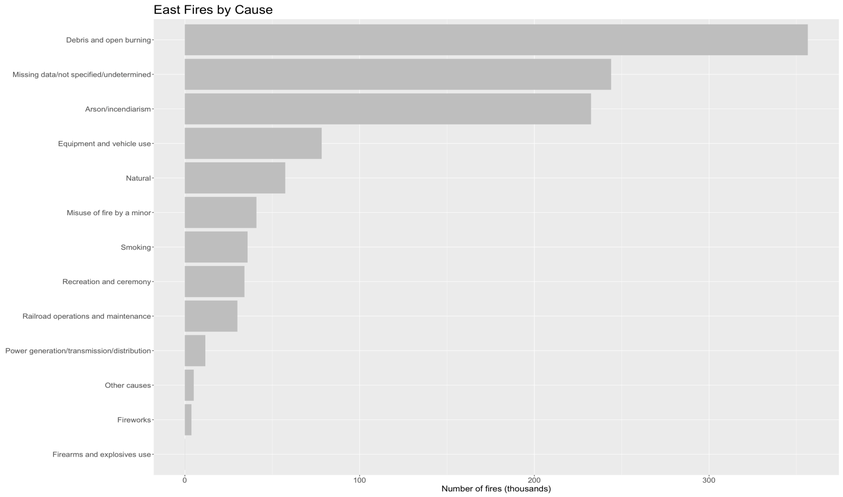


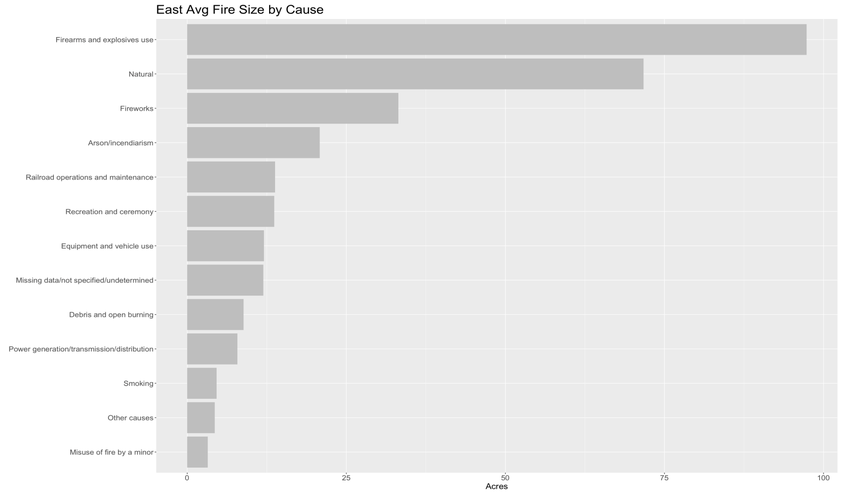
#Fires, Size and Days to Containment by Cause West & East US

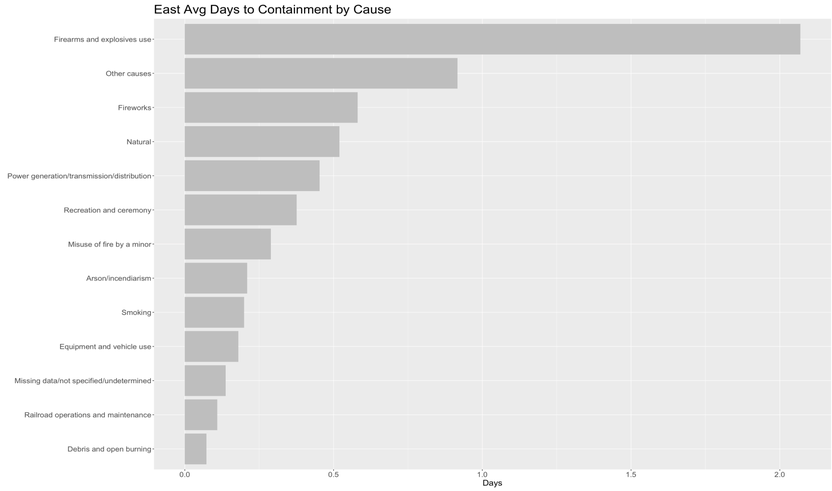




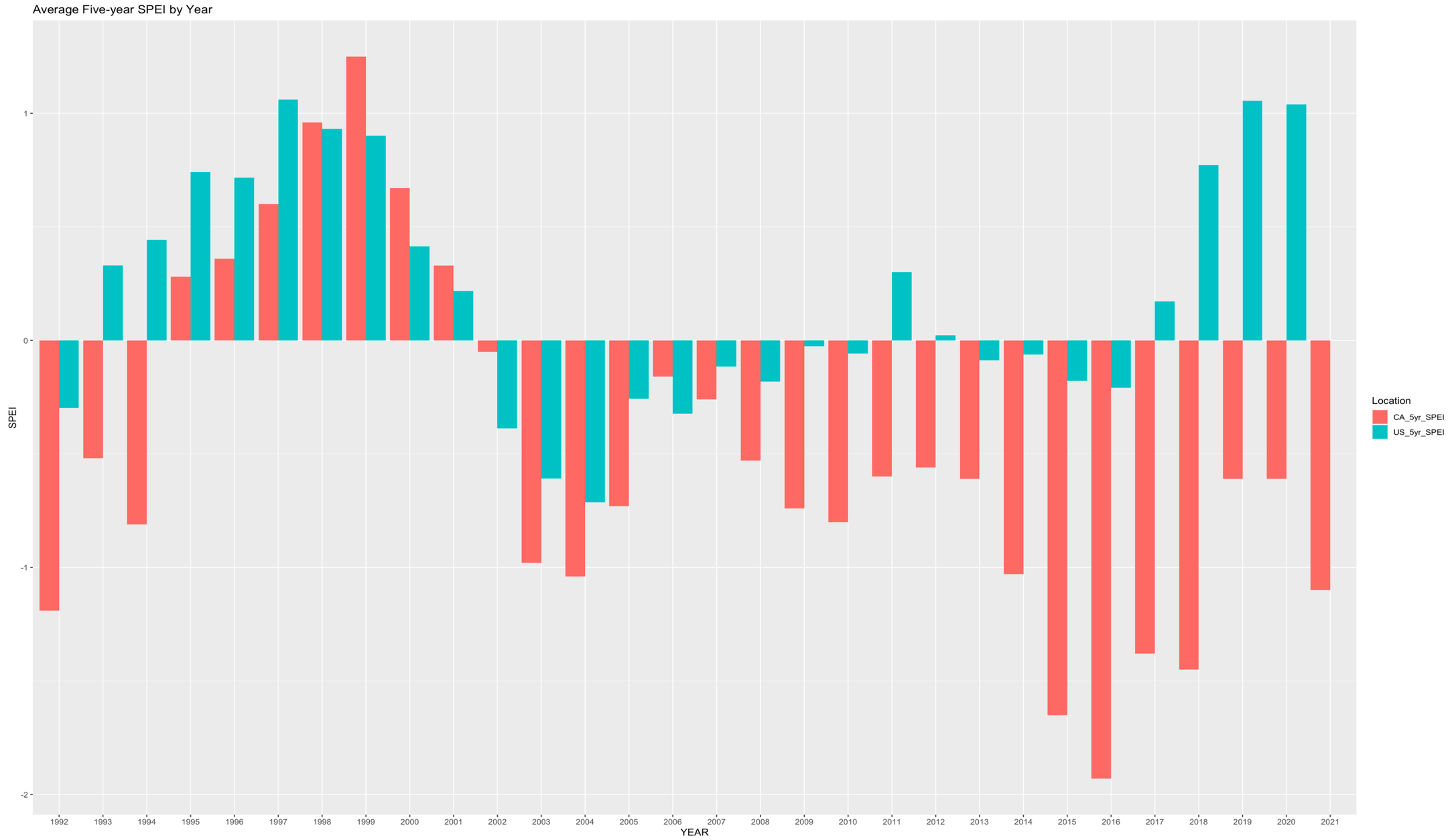








SPEI by Year US & CA



Correlation Matrix

Numeric Attributes

US

	DISCOVERY_TIME	CONT_TIME	FIRE_SIZE	LATITUDE	LONGITUDE	DAYS_TO_CONT
DISCOVERY_TIME	1.00	0.59	0.00	0.00	0.06	0.00
CONT_TIME	0.59	1.00	0.00	-0.03	0.03	-0.07
FIRE_SIZE	0.00	0.00	1.00	0.01	-0.03	0.16
LATITUDE	0.00	-0.03	0.01	1.00	-0.29	0.10
LONGITUDE	0.06	0.03	-0.03	-0.29	1.00	-0.11
DAYS_TO_CONT	0.00	-0.07	0.16	0.10	-0.11	1.00

CA

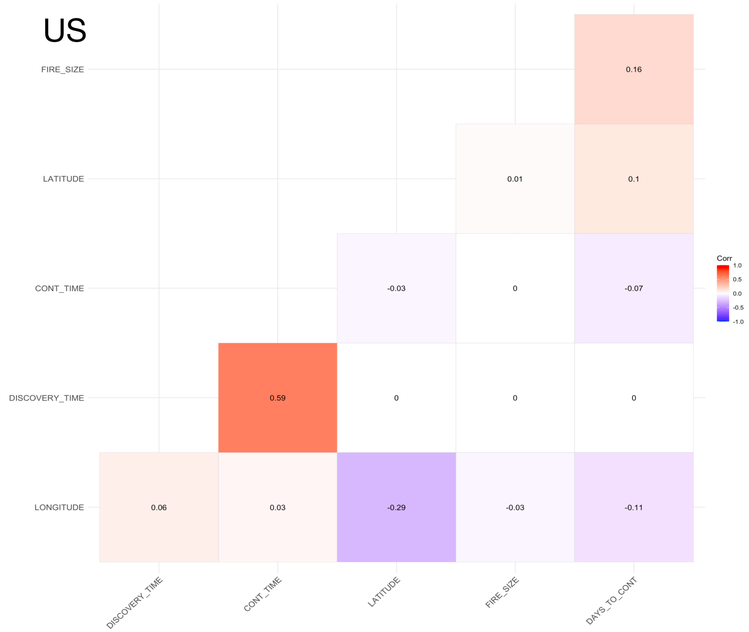
	DISCOVERY_TIME	CONT_TIME	FIRE_SIZE	LATITUDE	LONGITUDE	DAYS_TO_CONT
DISCOVERY_TIME	1.00	0.65	0.00	0.00	0.00	0.00
CONT_TIME	0.65	1.00	0.01	0.02	0.01	-0.02
FIRE_SIZE	0.00	0.01	1.00	0.00	0.00	0.22
LATITUDE	0.00	0.02	0.00	1.00	-0.88	0.05
LONGITUDE	0.00	0.01	0.00	-0.88	1.00	-0.02
DAYS_TO_CONT	0.00	-0.02	0.22	0.05	-0.02	1.00

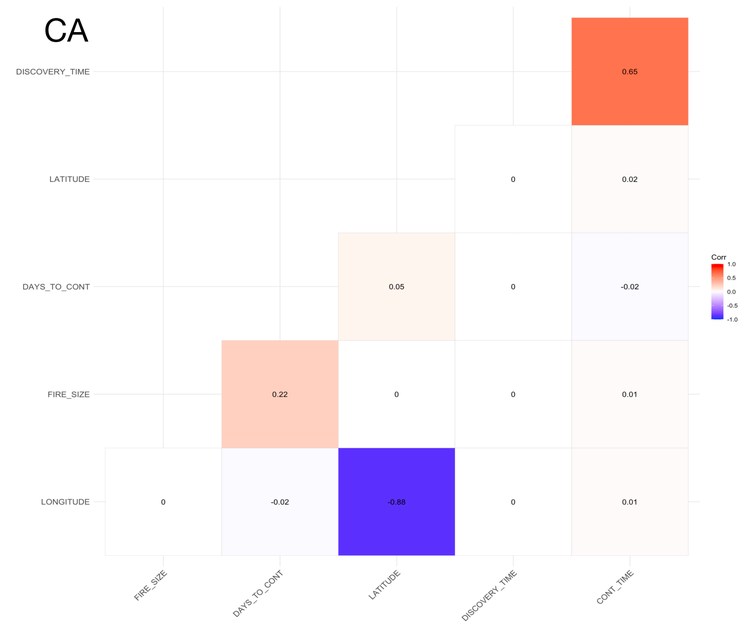
West

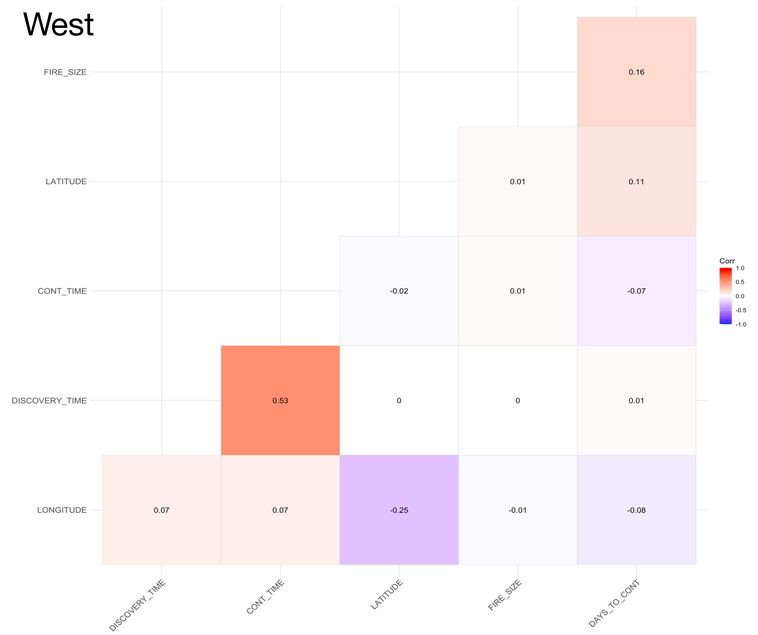
	DISCOVERY_TIME	CONT_TIME	FIRE_SIZE	LATITUDE	LONGITUDE	DAYS_TO_CONT
DISCOVERY_TIME	1.00	0.53	0.00	0.00	0.07	0.01
CONT_TIME	0.53	1.00	0.01	-0.02	0.07	-0.07
FIRE_SIZE	0.00	0.01	1.00	0.01	-0.01	0.16
LATITUDE	0.00	-0.02	0.01	1.00	-0.25	0.11
LONGITUDE	0.07	0.07	-0.01	-0.25	1.00	-0.08
DAYS_TO_CONT	0.01	-0.07	0.16	0.11	-0.08	1.00

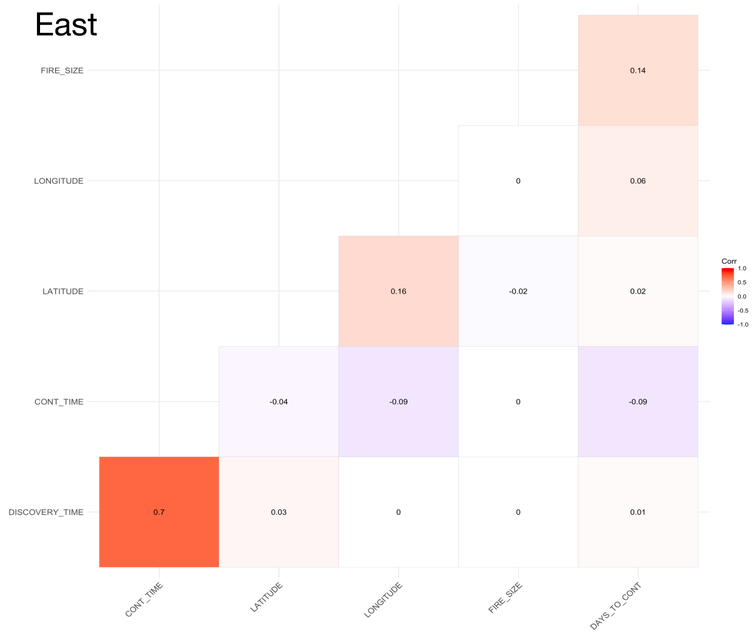
East

	DISCOVERY_TIME	CONT_TIME	FIRE_SIZE	LATITUDE	LONGITUDE	DAYS_TO_CONT
DISCOVERY_TIME	1.00	0.70	0.00	0.03	0.00	0.01
CONT_TIME	0.70	1.00	0.00	-0.04	-0.09	-0.09
FIRE_SIZE	0.00	0.00	1.00	-0.02	0.00	0.14
LATITUDE	0.03	-0.04	-0.02	1.00	0.16	0.02
LONGITUDE	0.00	-0.09	0.00	0.16	1.00	0.06
DAYS_TO_CONT	0.01	-0.09	0.14	0.02	0.06	1.00









- Fair positive correlation between DISCOVERY_TIME and CONT_TIME
- Very weak positive correlation between FIRE_SIZE and DAYS_TO_CONT
  - Possible that days is not granular enough to see real relationship since 0 - 23:59 hours is all 0 days.
  - Possible that external factors also have an impact like more resources are used to fight bigger fires

Linear Regression

CA Average Fire Size and SPEI

- H_0: $\beta = 0$
- H_A: $\beta < 0$

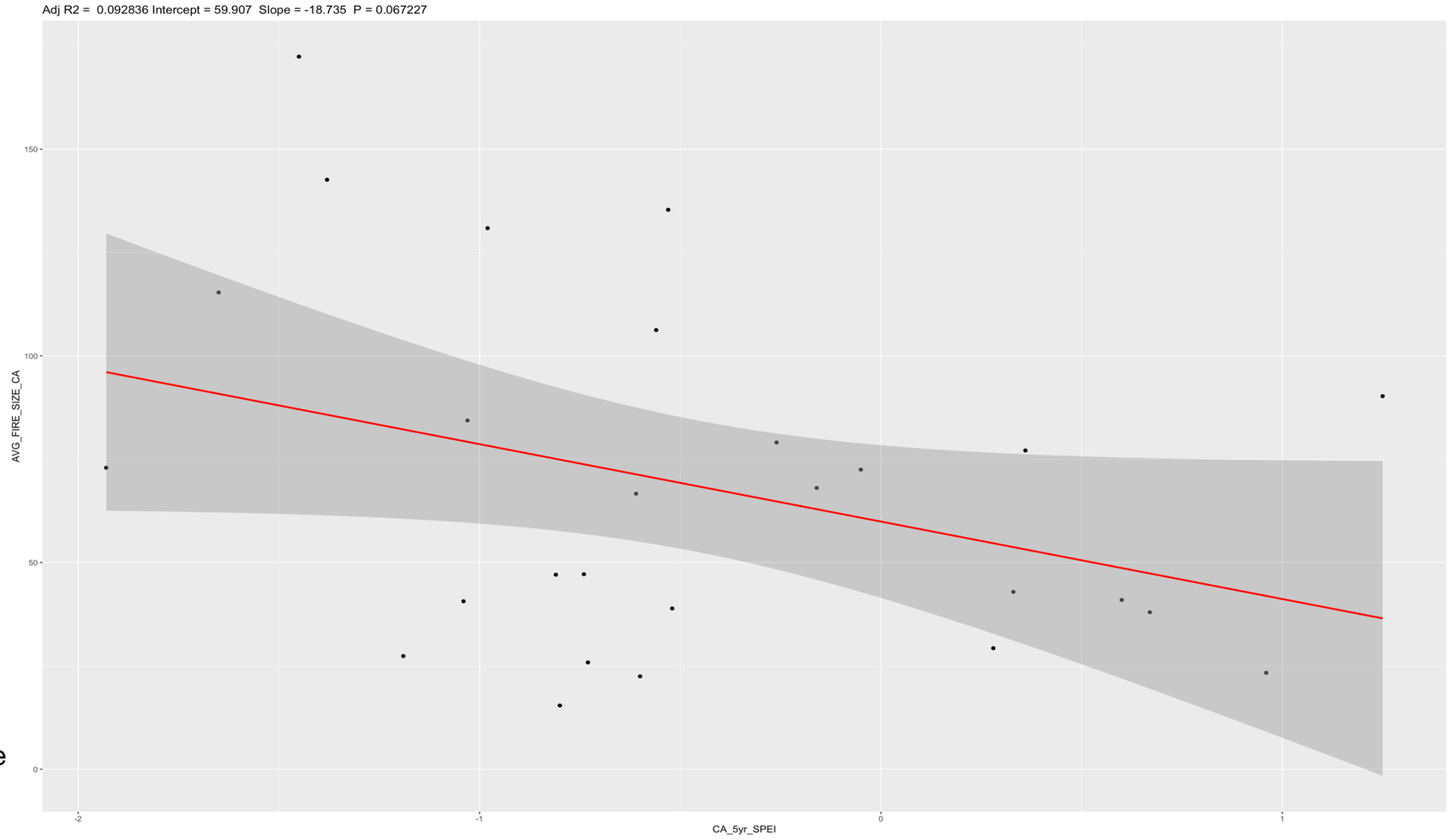
  …where β is an estimate of the slope of the regression line.

```
Call:
lm(formula = AVG_FIRE_SIZE_CA ~ CA_5yr_SPEI, data = ca_fs_spei)

Residuals:
   Min     1Q Median     3Q    Max
-59.49 -27.31  -7.73  24.24  85.33

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    59.91       8.97    6.68 0.00000053 ***
CA_5yr_SPEI   -18.74       9.79   -1.91      0.067 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40 on 25 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.128,     Adjusted R-squared:  0.0928
F-statistic: 3.66 on 1 and 25 DF,  p-value: 0.0672
```



- One-tailed test so P = 0.0335, we can reject H_0 and accept H_A at a 95% confidence level.
- Statistically significant relationship between average fire size and SPEI.

Linear Regression

CA Fire Frequency and SPEI

- H_0: $\beta = 0$
- H_A: $\beta < 0$

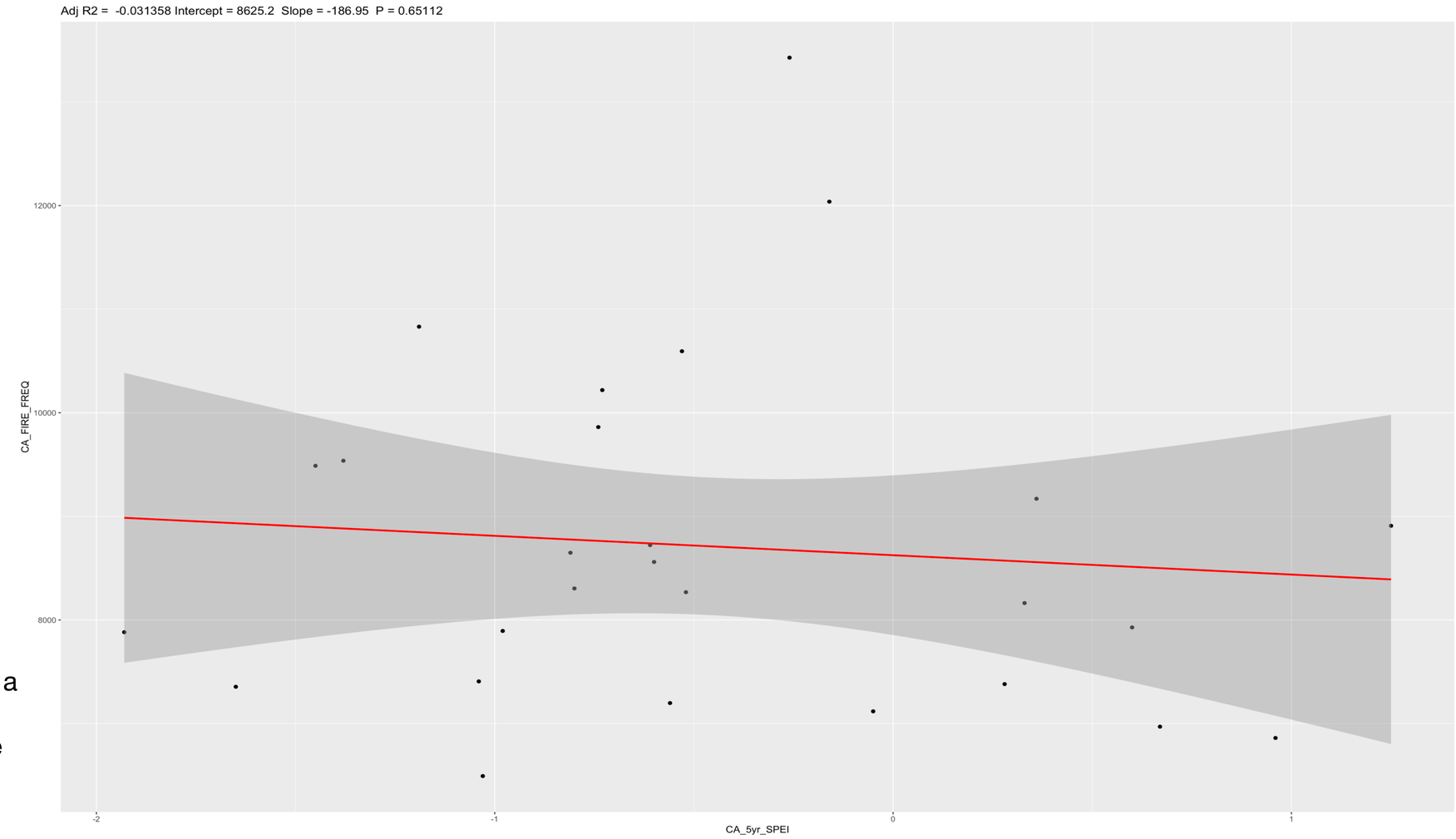
  …where β is an estimate of the slope of the regression line.

```
Call:
lm(formula = CA_FIRE_FREQ ~ CA_5yr_SPEI, data = ca_ff_spei)

Residuals:
   Min     1Q Median     3Q    Max
 -2325  -1302   -400    633   4754

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)     8625        374   23.05 <0.0000000000000002 ***
CA_5yr_SPEI     -187        408   -0.46                0.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1670 on 25 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.00831,   Adjusted R-squared:  -0.0314
F-statistic: 0.209 on 1 and 25 DF,  p-value: 0.651
```



- One-tailed test so P = 0.325, we fail to reject H_0 at a 95% confidence level.
- No statistically significant relationship between fire frequency and SPEI.

Linear Regression

CA Area Burned and SPEI

- H_0: $\beta = 0$
- H_A: $\beta < 0$

  …where β is an estimate of the slope of the regression line.

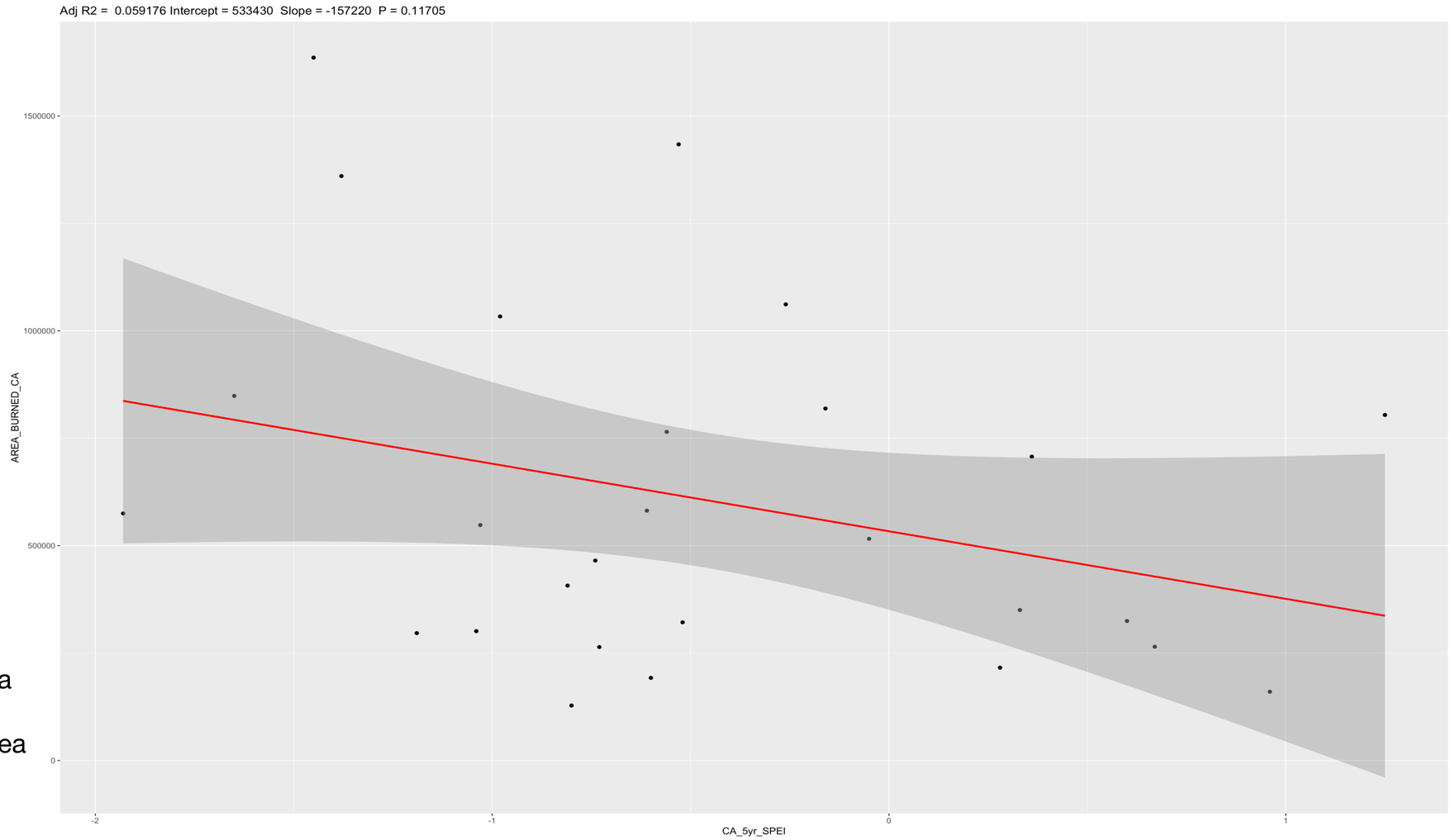
```
Call:
lm(formula = AREA_BURNED_CA ~ CA_5yr_SPEI, data = ca_fa_spei)

Residuals:
    Min      1Q  Median      3Q     Max
-531303 -267673 -131405  245401  874354

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   533429      88710    6.01 0.0000028 ***
CA_5yr_SPEI  -157224      96850   -1.62      0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 397000 on 25 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.0954,    Adjusted R-squared:  0.0592
F-statistic: 2.64 on 1 and 25 DF,  p-value: 0.117
```

- One-tailed test so P = 0.06, we fail to reject H_0 at a 95% confidence level.
- No statistically significant relationship between area burned and SPEI.



Linear Regression

CA Percentage of Large Fires and SPEI

(Large fire > 10,000 acres)

- H_0: $\beta = 0$
- H_A: $\beta < 0$

  …where β is an estimate of the slope of the regression line.

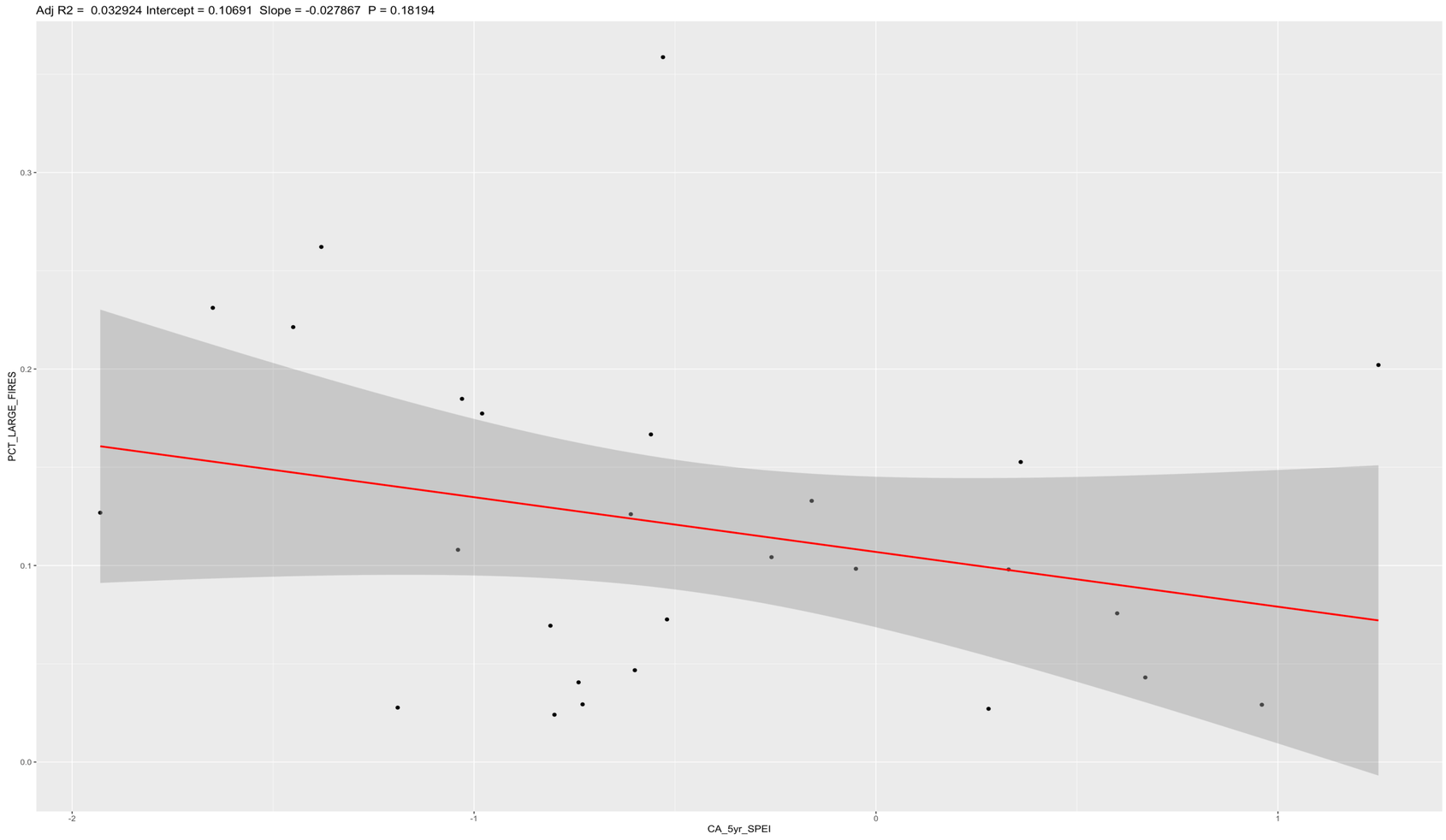
```
Call:
lm(formula = CA_PCT_LARGE_FIRES ~ CA_5yr_SPEI, data = ca_lf_spei)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11237 -0.05555 -0.00996  0.04670  0.23702

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.1069     0.0186    5.75 0.0000054 ***
CA_5yr_SPEI  -0.0279     0.0203   -1.37      0.18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.083 on 25 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.0701,    Adjusted R-squared:  0.0329
F-statistic: 1.89 on 1 and 25 DF,  p-value: 0.182
```

- One-tailed test so P = 0.09, we fail to reject H_0 at a 95% confidence level.
- No statistically significant relationship between percentage of large fires and SPEI.

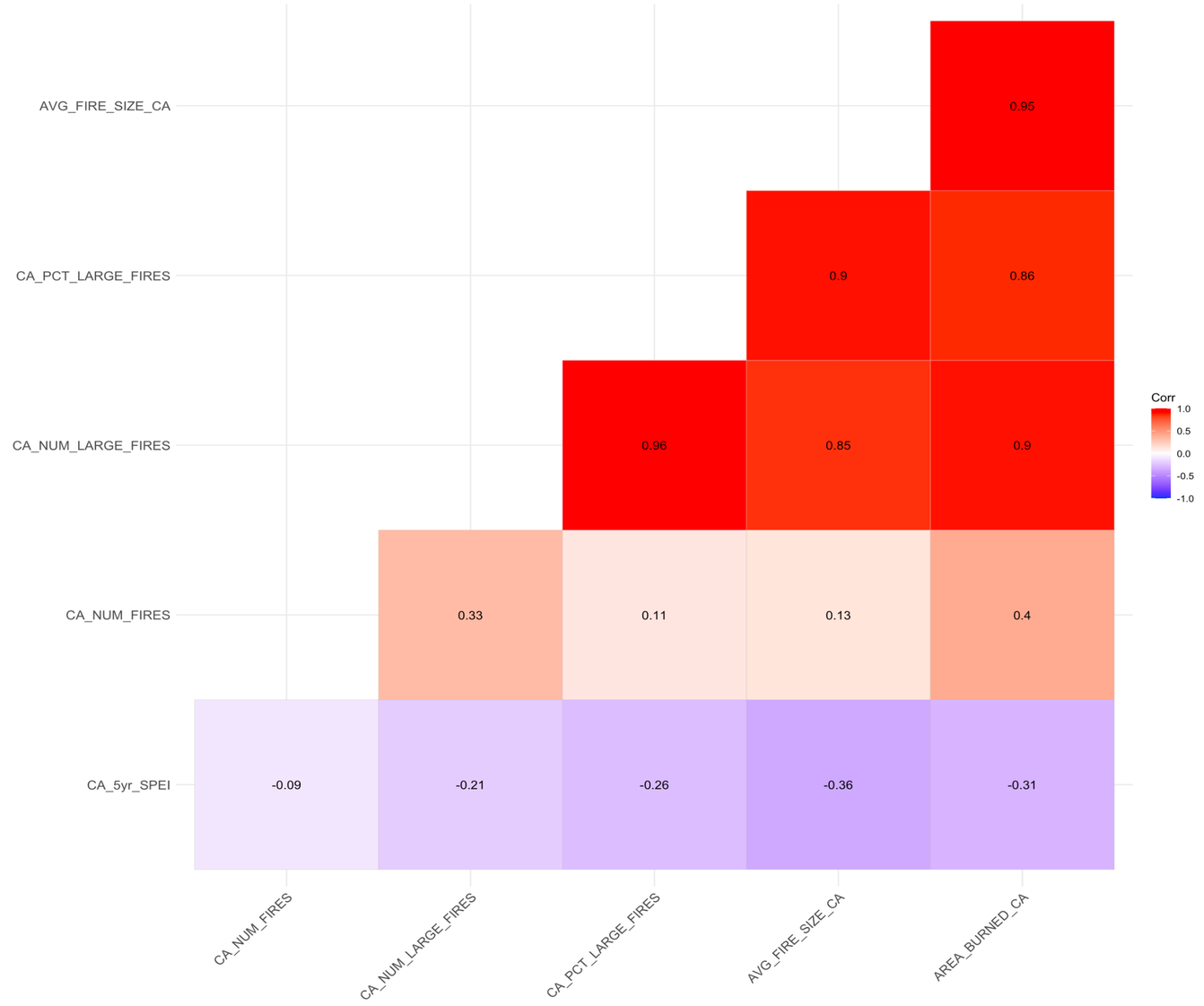


Correlation Matrix

CA Fire Size, Frequency, Area, Pct Large Fires

	AVG_FIRE_SIZE_CA	AREA_BURNED_CA	CA_NUM_FIRES	CA_NUM_LARGE_FIRES	CA_PCT_LARGE_FIRES	CA_5yr_SPEI
AVG_FIRE_SIZE_CA	1.00	0.95	0.13	0.85	0.90	-0.36
AREA_BURNED_CA	0.95	1.00	0.40	0.90	0.86	-0.31
CA_NUM_FIRES	0.13	0.40	1.00	0.33	0.11	-0.09
CA_NUM_LARGE_FIRES	0.85	0.90	0.33	1.00	0.96	-0.21
CA_PCT_LARGE_FIRES	0.90	0.86	0.11	0.96	1.00	-0.26
CA_5yr_SPEI	-0.36	-0.31	-0.09	-0.21	-0.26	1.00

- SPEI most strongly correlated with average fire size which supports the Linear regression model conclusions which show that SPEI and average fire size have a statistically significant relationship.



Linear Regression Prediction

CA AVG_FIRE_SIZE Prediction

- Since the wildfire data only goes up to 2018 I would like to predict the average fire size in 2019, 2020 and 2021 base on SPEI values for those years.
- ca_fs_spei dataframe below with NA values for 2019-2021:

Year <fctr>	AVG_FIRE_SIZE_CA <dbl>	CA_5yr_SPEI <dbl>
1992	27	−1.19
1993	39	−0.52
1994	47	−0.81
1995	29	0.28
1996	77	0.36
1997	41	0.60
1998	23	0.96
1999	90	1.25
2000	38	0.67
2001	43	0.33
2002	72	−0.05
2003	131	−0.98
2004	41	−1.04
2005	26	−0.73
2006	68	−0.16
2007	79	−0.26
2008	135	−0.53
2009	47	−0.74
2010	15	−0.80
2011	22	−0.60
2012	106	−0.56
2013	67	−0.61
2014	84	−1.03
2015	115	−1.65
2016	73	−1.93
2017	143	−1.38
2018	172	−1.45
2019	*NA*	−0.61
2020	*NA*	−0.61
2021	*NA*	−1.10

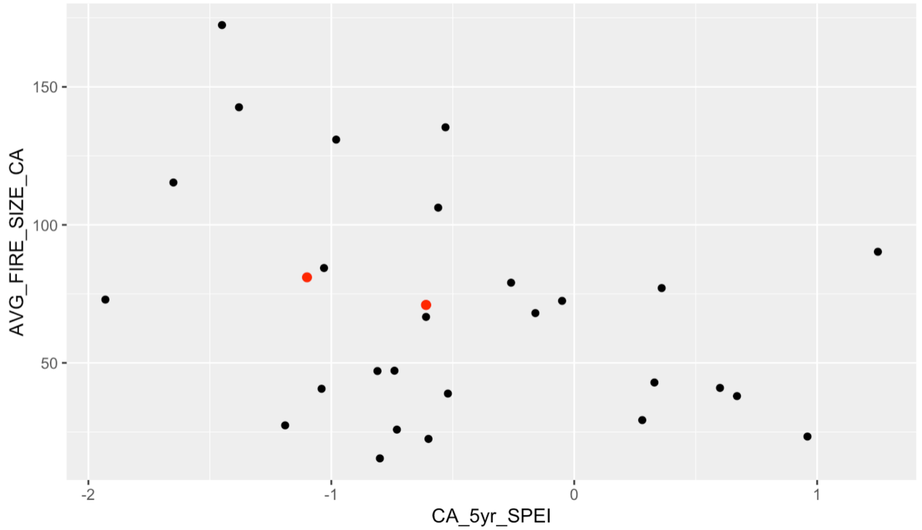
```
ca_SPEI_new <- data.frame(CA_5yr_SPEI=c(-0.61, -0.61, -1.10))
ca_avg_fs_pred <- data.frame(Year = c(2019, 2020, 2021), AVG_FIRE_SIZE_CA = predict(fit1, ca_SPEI_new))
ca_avg_fs_pred
```

	Year <dbl>	AVG_FIRE_SIZE_CA <dbl>
1	2019	71
2	2020	71
3	2021	81

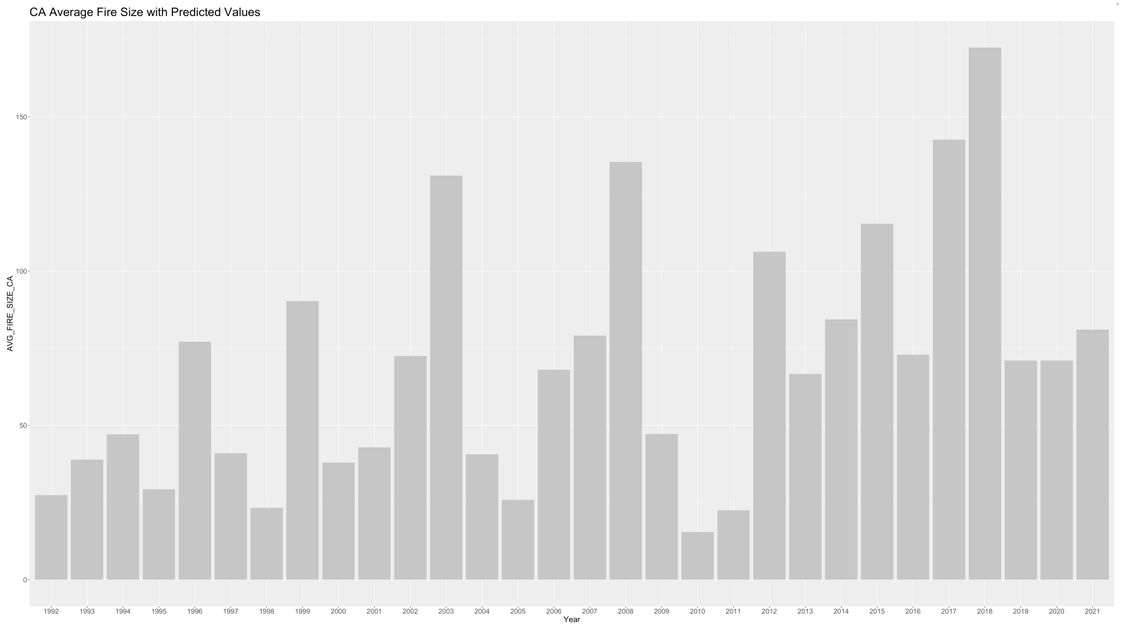
Updated Dataframe with new predicted values:

Year <fctr>	AVG_FIRE_SIZE_CA <dbl>	CA_5yr_SPEI <dbl>
1992	27	−1.19
1993	39	−0.52
1994	47	−0.81
1995	29	0.28
1996	77	0.36
1997	41	0.60
1998	23	0.96
1999	90	1.25
2000	38	0.67
2001	43	0.33
2002	72	−0.05
2003	131	−0.98
2004	41	−1.04
2005	26	−0.73
2006	68	−0.16
2007	79	−0.26
2008	135	−0.53
2009	47	−0.74
2010	15	−0.80
2011	22	−0.60
2012	106	−0.56
2013	67	−0.61
2014	84	−1.03
2015	115	−1.65
2016	73	−1.93
2017	143	−1.38
2018	172	−1.45
2019	71	−0.61
2020	71	−0.61
2021	81	−1.10

Scatter Plot with Predicted Average Fire Size Values in Red



CA Average Fire Size with Predicted Values



Conclusion

- Data was messy and difficult to manipulate with a lot of time spent cleaning and transforming. As such opportunities were missed for further insight and analysis.
- Fires seem to be getting worse in Western US states where it’s getting drier.
- Average fire size and SPEI for CA showed a statistically significant relationship.
  - This enabled me to predict the average fire size in CA for 2019-2021 based on the latest SPEI values.
- I think hours to containment may show a stronger relationship with SPEI than days to containment if it was calculated and modeled due to the number of fires contained within a day.
- Further work could look at modeling the categorical variables particularly fire causes.

Appendix

Complete Fires Attribute Information

new_fires attributes in green
dropped attributes in black

Fires:
Table including wildfire data for the period of 1992-2018 compiled from US federal, state, and local reporting systems.

FOD_ID = Unique numeric record identifier.

FPA_ID = Unique identifier that contains information necessary to track back to the original record in the source dataset.

SOURCE_SYSTEM_TYPE = Type of source database or system that the record was drawn from (federal, nonfederal, or interagency).

SOURCE_SYSTEM = Name of or other identifier for source database or system that the record was drawn from. See Table 1 in Short (2014), or \Supplements\FPA_FOD_source_list.pdf, for a list of sources and their identifier.

NWCG_REPORTING_AGENCY = Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).

NWCG_REPORTING_UNIT_ID = Active NWCG Unit Identifier for the unit preparing the fire report.

NWCG_REPORTING_UNIT_NAME = Active NWCG Unit Name for the unit preparing the fire report.

SOURCE_REPORTING_UNIT = Code for the agency unit preparing the fire report, based on code/name in the source dataset.

SOURCE_REPORTING_UNIT_NAME = Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.

LOCAL_FIRE_REPORT_ID = Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.

LOCAL_INCIDENT_ID = Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.

FIRE_CODE = Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/).

FIRE_NAME = Name of the incident, from the fire report (primary) or ICS-209 report (secondary).

ICS_209_PLUS_INCIDENT_JOIN_ID = Primary identifier needed to join into operational situation reporting data for the incident in the ICS-209-PLUS dataset.

ICS_209_PLUS_COMPLEX_JOIN_ID = If part of a complex, secondary identifier potentially needed to join to operational situation reporting data for the incident in the ICS-209-PLUS dataset (2014 and later only).

MTBS_ID = Incident identifier, from the MTBS perimeter dataset.

MTBS_FIRE_NAME = Name of the incident, from the MTBS perimeter dataset.

COMPLEX_NAME = Name of the complex under which the fire was ultimately managed, when discernible.

FIRE_YEAR = Calendar year in which the fire was discovered or confirmed to exist.

DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.

DISCOVERY_DOY = Day of year on which the fire was discovered or confirmed to exist.

DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.

NWCG_CAUSE_CLASSIFICATION = Broad classification of the reason the fire occurred (Human, Natural, Missing data/not specified/undetermined).

NWCG_GENERAL_CAUSE = Event or circumstance that started a fire or set the stage for its occurrence (Arson/incendiarism, Debris and open burning, Equipment and vehicle use, Firearms and explosives use, Fireworks, Misuse of fire by a minor, Natural, Power generation/transmission/distribution, Railroad operations and maintenance, Recreation and ceremony, Smoking, Other causes, Missing data/not specified/undetermined).

NWCG_CAUSE_AGE_CATEGORY = If cause attributed to children (ages 0-12) or adolescents (13-17), the value for this data element is set to Minor; otherwise null.

CONT_DATE = Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).

CONT_DOY = Day of year on which the fire was declared contained or otherwise controlled.

CONT_TIME = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).

FIRE_SIZE = The estimate of acres within the final perimeter of the fire.

FIRE_SIZE_CLASS = Code for fire size based on the number of acres within the final fire perimeter (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).

LATITUDE = Latitude (NAD83) for point location of the fire (decimal degrees).

LONGITUDE = Longitude (NAD83) for point location of the fire (decimal degrees).

OWNER_DESCR = Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.

STATE = Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.

COUNTY = County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.

FIPS_CODE = Five-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.

FIPS_NAME = County name from the FIPS publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.

List of R libraries used

- library(RSQLite)
- library(dbplyr)
- library(dplyr)
- library(tidyr)
- library(ggthemes) #for visual themes
- library(lubridate) #for date conversion
- library(chron) #for time conversion
- library(magrittr) #call and update with %<>%
- library(pastecs) #descriptive stats
- library(ggplot2) #for visuals
- library(mosaicData) #for correlation matrix
- library(ggcorrplot) #for linear regression
- library(scales) #for normalizing
- library(ggpubr) #for ggarrange
- library(viridis) #for color scale
- library(hrbrthemes) #themes for ggplot2

Assignment Instructions

1.The dataset should have a minimum of 50,000 to 100,000 Records(100,000 Preferred)

2.Provide Description about your dataset

3.Describe the variables used in your dataset

4.Provide the summary of the dataset (if you are using more than one dataset for comparison, provide the summary for both the datasets)

5.Provide the Descriptive Statistics for each variable

a)Min, Max, sum, range, mean, median, var, std. dev, NULLs, NAs(if any)

6.Provide the Descriptive Statistics of the dataset by:

a)Histogram

b)Box Plot

c)Pie Chart

d)Any other visualization

7.Find the overall mean for each variable

8.Predictive Modeling.

a)Use Linear Regression in two different models (You can use any models that can be applied to your data)

9.Conclusion of your project

The points(30) will be given based on the presentation, code submission and the output.

During the final class on June 12th, you will be presenting the final draft of your presentation.

While submitting the final project, please include the following:

- Final presentation (PPTX or PDF or Word)
- .RMD File
- & HTML File output