Soil data analysis Jupyter Notebook

The following is the Jupyter Notebook that was used to analyze soil data for the manuscript "PFAS soil and groundwater contamination via industrial airborne emission and land deposition in SW Vermont and Eastern New York State" by Tim Schroeder, David Bond, and Janet Foley. All code executed and graphs are included in this file.

```
In [116]: #Import required packages
          import os
          import pandas as pd
          import altair as alt
          import scipy.stats as stats
          #import researchpy as rp
          import statsmodels.api as sm
          from statsmodels.formula.api import ols
          import matplotlib.pyplot as plt
```

Soil QA/QC analyses

In order to test the possibility that cross-contamination from sampling equipment may have contributed to the pattern of PFAS in soils observed in this study, we plot the concentration of PFOA and PFOS versus the sequence order in which samples were collected, with samples color-coded based on the region of the sample site.

In [117]:

```
#set working directory, read datafile, and inspect dataframe
os.chdir("/Users/tschroeder/Jupyter_Files/Soil_PFOA")
df = pd.read_csv("BC_soils_orderedbydate_2.csv")
df.head(10)
```

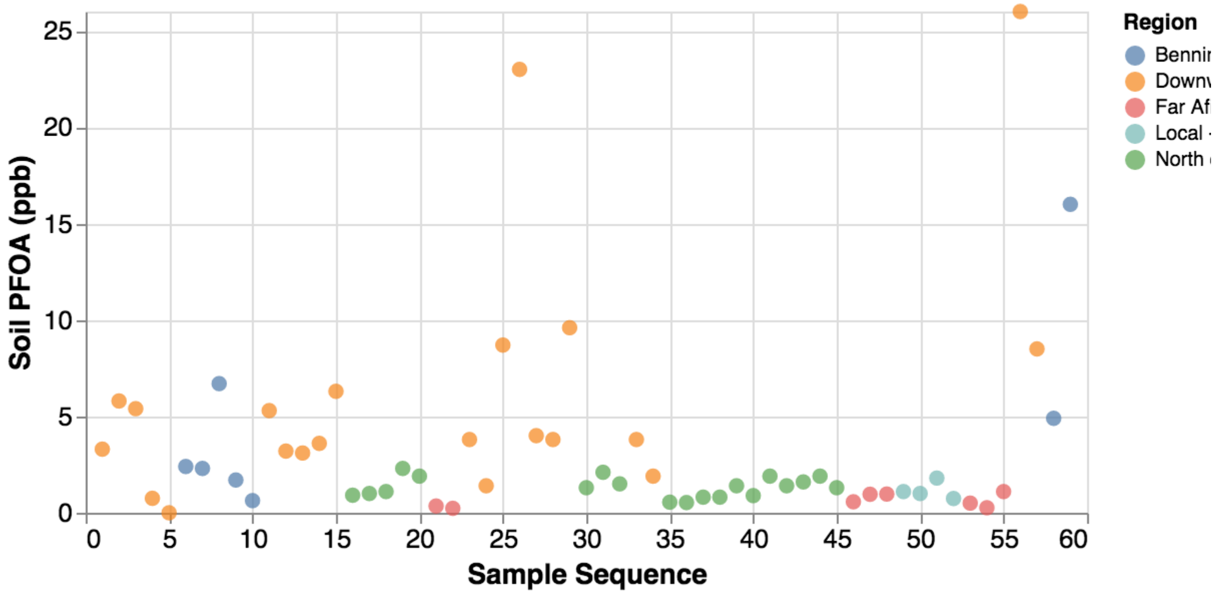
Out[117]:

	Sequence	Sample #	Region	Sample name	Easting	Northing	PFOA	PFOS	PFHpA	PFHxA
0	1	7-20-17-1S	Downwind of Bennington	BaldMtnWest-2017#1	649307	4750384	3.30	NaN	0.27	0.26
1	2	7-20-17-2S	Downwind of Bennington	BaldMtnWest-2017#2	649826	4751193	5.80	0.46	0.44	0.24
2	3	7-20-17-3S	Downwind of Bennington	BaldMtnWest-2017#3	650314	4752322	5.40	0.35	0.41	0.22
3	4	7-20-17-4S	Downwind of Bennington	BaldMtnWest-2017#4	650504	4752550	0.75	0.34	NaN	NaN
4	5	7-20-17-5S	Downwind of Bennington	BaldMtnWest-2017#5	651460	4753661	0.00	NaN	NaN	NaN
5	6	7-21-17-1S	Bennington Local	Honeysuckle-lane	644521	4753676	2.40	0.42	NaN	NaN
6	7	7-21-17-2S	Bennington Local	Matteson Rd	644813	4753972	2.30	NaN	0.32	0.16
7	8	7-21-17-3S	Bennington Local	Rice Lane	645245	4753113	6.70	NaN	0.23	0.13
8	9	7-24-17-1S	Bennington Local	Rt. 7a	646189	4752348	1.70	NaN	NaN	0.15
9	10	7-24-17-2S	Bennington Local	Chapel Rd.	648608	4751518	0.63	NaN	0.41	NaN

In [118]:

```
#create plot of sample sequence vs. PFOA concentration
ytitle="Soil PFOA (ppb)"
xtitle="Sample Sequence"
alt.Chart(df).mark_circle(size=60).encode(
    alt.X('Sequence', title=xtitle),
    alt.Y('PFOA',title = ytitle, scale=alt.Scale(
            domain=(0, 25),
            clamp=True)),
    color='Region',
    tooltip=['Sample #', 'Region', 'PFOA', 'PFOS']).properties(width=500
, height=250).configure_axis(labelFontSize=13,
    titleFontSize=14)
```

Out[118]:

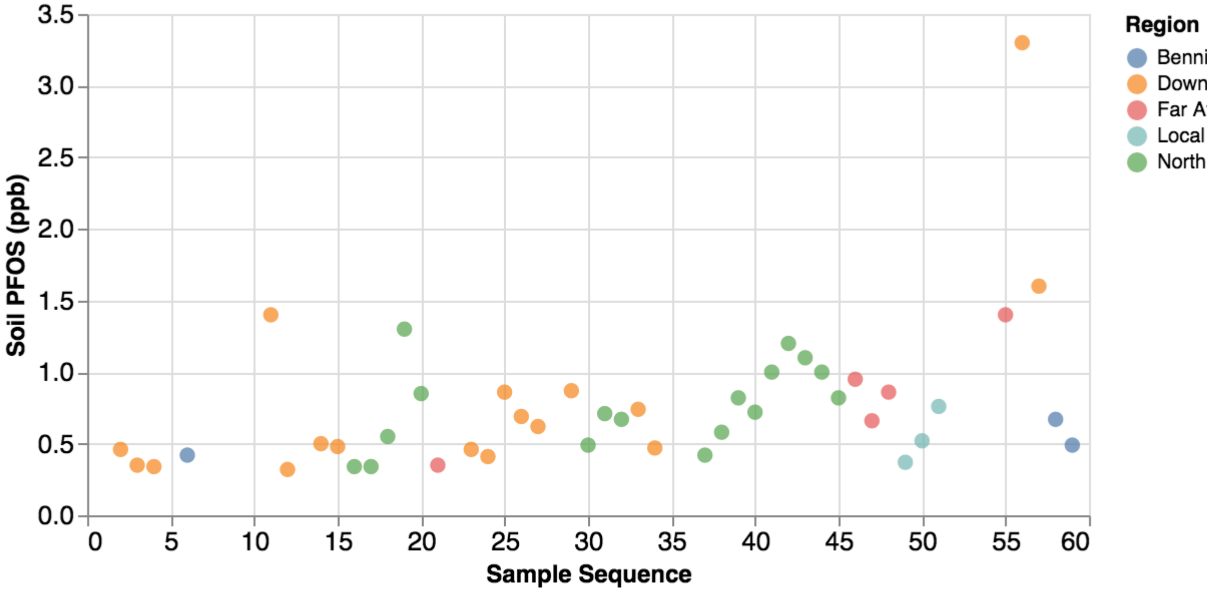


SI Figure 1: Sample sequence vs. PFOA Concentration It can be seen qualitatively that soil sample PFOA concentration has a higher dependence on the region from which the sample is from than the sample that immediately precedes it. There is high variability in the samples from the impacted region (Bennington Local and Downwind of Bennington). When we moved from sampling in an impacted region to one of the theoretically non-impacted regions, the first sample's PFOA concentration in the non-impacted region was always consistent with the median value in that region. The two outlier samples (#26 - 23 ppt PFOA, and #56 - 98 ppt PFOA) do not appear to have an impact on the concentration in the samples collected after these.

In [119]:

```
ytitle="Soil PFOS (ppb)"
xtitle="Sample Sequence"
alt.Chart(df).mark_circle(size=60).encode(
    alt.X('Sequence', title=xtitle),
    alt.Y('PFOS',title = ytitle, scale=alt.Scale(
            domain=(0, 3.5),
            clamp=True)),
    color='Region',
    tooltip=['Sample #', 'Region', 'PFOA', 'PFOS']).properties(width=500
, height=250).configure_axis(labelFontSize=13,
    titleFontSize=12)
```

Out[119]:



SI Figure 2: Sample sequence vs. PFOS Concentration Soil PFOS concentration, which is present at statistically similar levels in soil in all five sampling regions, does not appear to depend in any sample on prior-collected samples. The outlier (#56 - 3.3 ppt PFOS) is followed by a sample with higher than average PFOS.

Comparison of Multiple Soil Studies

The next test that we ran on our soil dataset was to compare the results with other soil PFAS studies conducted in the area in the same timeframe. These include:

- Soil samples collected by the Vermont Dept. of Environmental Conservation (VT-DEC) around the North Bennington area impacted by the ChemFab factory contamination (https://dec.vermont.gov/commissioners-office/pfoa (https://dec.vermont.gov/commissioners-office/pfoa))
- Soil samples collected for preparation of the Draft Conceptual Site Model Site Investigation Report prepared by Barr Engineering on behalf of St. Gobain Performance Plastics; samples collected around the Bennington region impacted by ChemFab (https://anrweb.vt.gov/PubDocs/DEC/PFOA/Conceptual%20Site%20Model%20Site%20Investigation/DRAF
CSM-Site-Investigation-Report-text-only-FEB2018.pdf
(https://anrweb.vt.gov/PubDocs/DEC/PFOA/Conceptual%20Site%20Model%20Site%20Investigation/DRAF
CSM-Site-Investigation-Report-text-only-FEB2018.pdf))
- Samples collected in a forested region of Bennington by a contractor for a solar developer as part of the permitting process for a solar farm on the site (https://epuc.vermont.gov/?q=node/64/127312/FV-PFEXAFF-PTL (https://epuc.vermont.gov/?q=node/64/127312/FV-PFEXAFF-PTL))
- PFAS soil background study across Vermont commissioned by VT-DEC (https://anrweb.vt.gov/PubDocs/DEC/PFOA/Soil-Background/PFAS-Background-Vermont-Shallow-Soils-03-24-19.pdf (https://anrweb.vt.gov/PubDocs/DEC/PFOA/Soil-Background/PFAS-Background-Vermont-Shallow-Soils-03-24-19.pdf)).

For comparison basis, the samples from this study are divided into those collected in areas hypothesized to be impacted by air emission from manufacturers (Bennington Local and Downwind), and those hypothesized to be not impacted (i.e. peripheral: North of Wind Pattern, Upwind, and Far Afield).

In [120]:

```
#Import datafile for the multiple studies comparrison
df = pd.read_csv("complete_soils_data_origins&cover_16Jan_4.csv")
df.head(10)
```

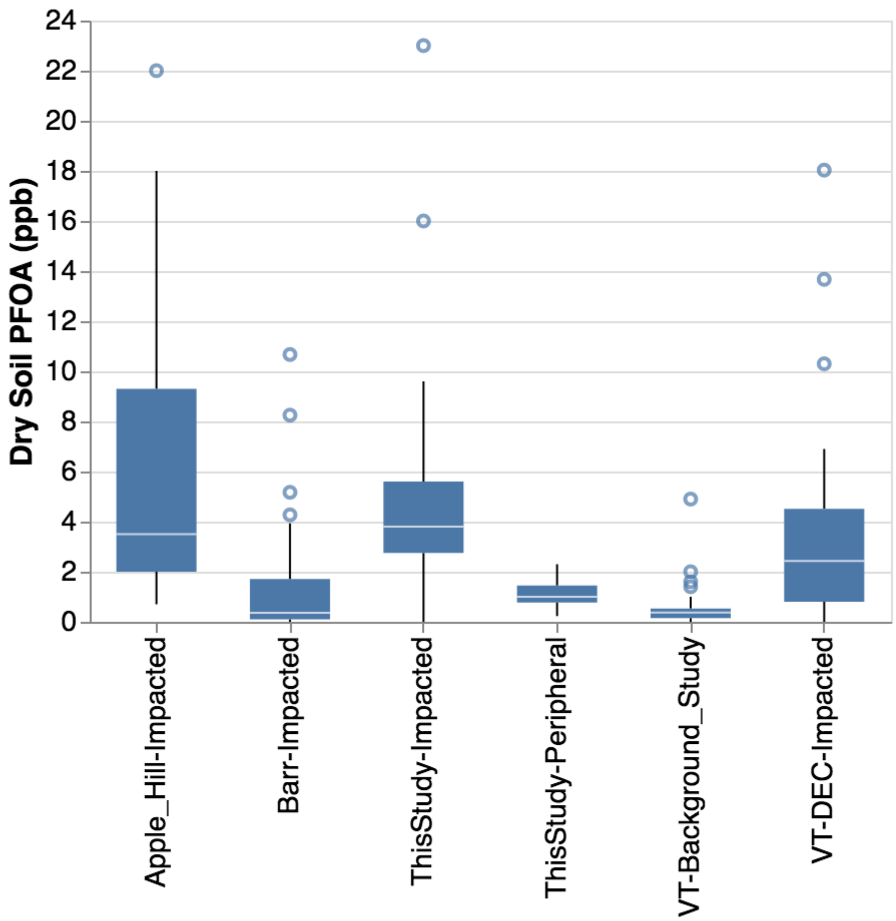
Out[120]:

	Source	LandCover	PFOA	PFOS
0	ThisStudy-Impacted	Grass/Pasture	3.58	4.25
1	ThisStudy-Impacted	Grass/Pasture	6.54	0.61
2	ThisStudy-Impacted	Grass/Pasture	1.53	0.18
3	ThisStudy-Impacted	Grass/Pasture	4.60	0.18
4	ThisStudy-Impacted	Developed	2.40	0.42
5	ThisStudy-Impacted	Developed	2.30	0.00
6	ThisStudy-Impacted	Developed	6.70	0.00
7	ThisStudy-Impacted	Developed	1.70	0.00
8	ThisStudy-Impacted	Developed	0.63	0.00
9	ThisStudy-Impacted	Developed	3.40	0.00

In [121]:

```
#create a boxplot based on the grouping of the samples by their source study
ytitle=" Dry Soil PFOA (ppb)"
xtitle=""
alt.Chart(df).mark_boxplot(size=40).encode(
    x=alt.X('Source', title=xtitle),
    y=alt.Y('PFOA', title=ytitle)).properties(width=400, height=300).configure_axis(labelFontSize=14,
    titleFontSize=15)
```

Out[121]:



SI Figure 3: Boxplots of soil PFOA Concentrations from Multiple Studies

Soil PFOA Statistical Analysis

Below, we run an ANOVA and Tukey pairwise analysis of soil PFOA concentration from the six study divisions to test for statistical difference bewtten them.

In [122]:

```
#run one-way ANOVA test on the sample groups
lm = ols('PFOA ~ Source',data=df).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
             df       sum_sq     mean_sq          F        PR(>F)
Source      5.0   796.507029  159.301406  17.201857  2.451483e-14
Residual  218.0  2018.834755    9.260710        NaN           NaN
```

The p value of 2.45e-14 is less than 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

In [123]:

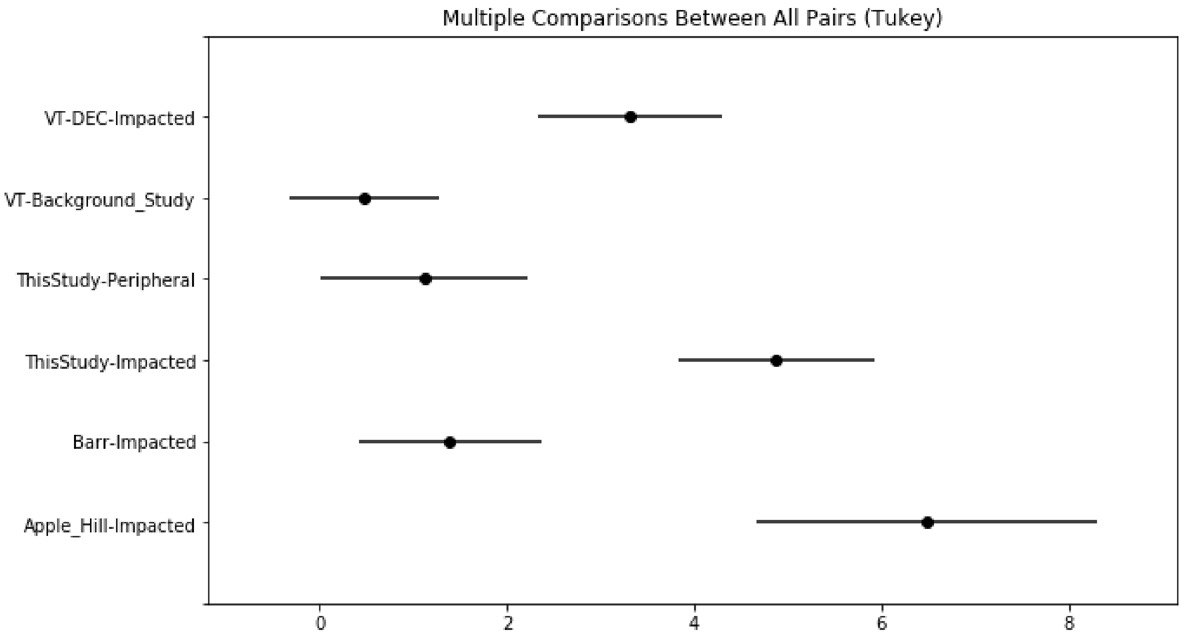
```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
#Run the Tukey analysis on all six sample groupings for PFOA
tukey = pairwise_tukeyhsd(endog=df['PFOA'],     # Data
                          groups=df['Source'],   # Groups
                          alpha=0.05)           # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[123]:

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05

              group1               group2  meandiff   p-adj    lower    upper  reject
  Apple_Hill-Impacted        Barr-Impacted   -5.0889   0.001  -7.8823  -2.2955    True
  Apple_Hill-Impacted   ThisStudy-Impacted   -1.6038  0.5729  -4.4457   1.2382   False
  Apple_Hill-Impacted ThisStudy-Peripheral   -5.3627   0.001  -8.2538  -2.4715    True
  Apple_Hill-Impacted  VT-Background_Study    -6.011   0.001   -8.666   -3.356    True
  Apple_Hill-Impacted      VT-DEC-Impacted   -3.1692  0.0165  -5.9714  -0.3671    True
        Barr-Impacted   ThisStudy-Impacted    3.4851   0.001   1.4599   5.5103    True
        Barr-Impacted ThisStudy-Peripheral   -0.2738     0.9  -2.3675   1.8199   False
        Barr-Impacted  VT-Background_Study   -0.9221  0.6362  -2.6754   0.8311   False
        Barr-Impacted      VT-DEC-Impacted    1.9196  0.0608  -0.0494   3.8886   False
   ThisStudy-Impacted ThisStudy-Peripheral   -3.7589   0.001  -5.9169  -1.6009    True
   ThisStudy-Impacted  VT-Background_Study   -4.4072   0.001  -6.2368  -2.5776    True
   ThisStudy-Impacted      VT-DEC-Impacted   -1.5655  0.2375  -3.6027   0.4718   False
 ThisStudy-Peripheral  VT-Background_Study   -0.6483     0.9  -2.5534   1.2568   False
 ThisStudy-Peripheral      VT-DEC-Impacted    2.1934  0.0357   0.0881   4.2988    True
  VT-Background_Study      VT-DEC-Impacted    2.8417   0.001   1.0745   4.6089    True
```



SI Figure 4: Soil PFOA Tukey variance overlap between studies: This analysis indicates statistical similarity between three of the four studies conducted in the impacted area, including the samples analyzed in this study. The samples collected by Barr Engineering from the impacted area are statistically similar to the two groups of samples from the not-impacted areas, and also to the VT-DEC samples. The lower concentration of PFOA in the Barr samples may be related to land cover in their collection locations. This relationship is explored in greater depth below.

Re-run PFOA soil analysis with peripheral groups removed

Below, we re-run the tukey comparison with the two groups of peripheral samples removed. This allows a smaller alpha value for each individual dataset to be used for each group in the comparison to arrive at a alpha of 0.05 for the analysis, which results in a lower threshold for statistical similarity.

In [124]:

```python
#remove the two peripheral studies from the dataframe
df2 = df[df['Source'] != 'VT-Background_Study']
df2 = df2[df2['Source'] != 'ThisStudy-Peripheral']
```

In [125]:

```python
#run one-way ANOVA test on the sample groups (without the peripheral area groups)
lm = ols('PFOA ~ Source',data=df2).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
             df        sum_sq     mean_sq         F    PR(>F)
Source      3.0    360.842300  120.280767  7.468442  0.000124
Residual  123.0   1980.939944   16.105203       NaN       NaN
```

The p value of 0.000124 is less than 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

In [126]:

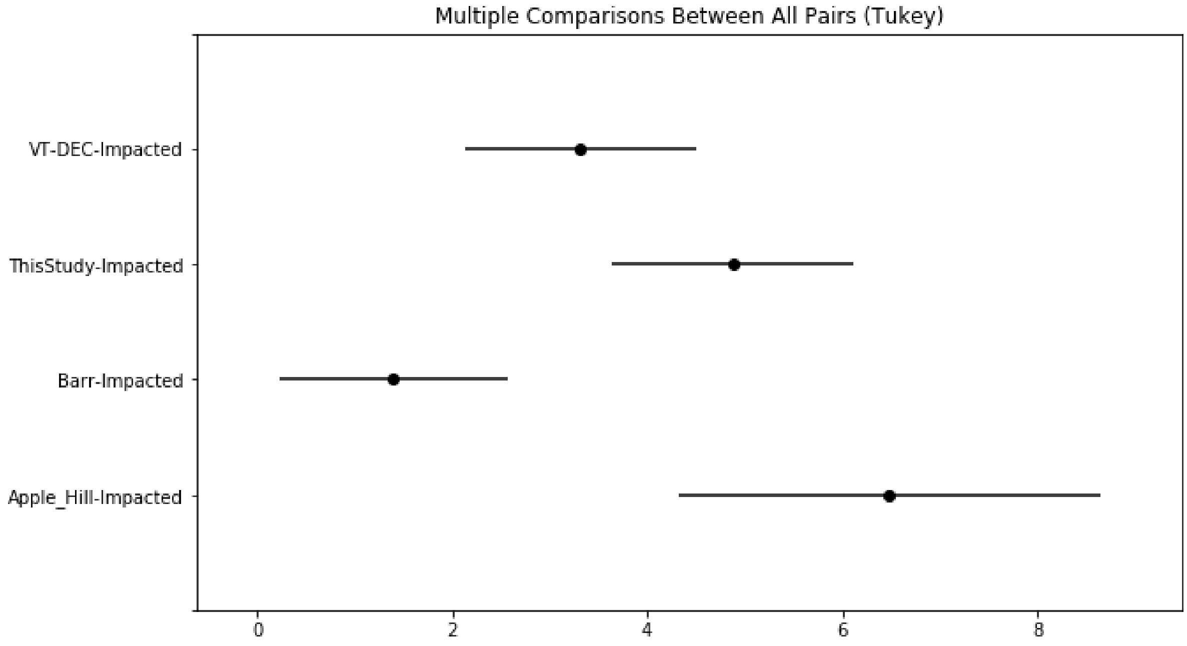
```
#Run the Tukey analysis on the four sample groupings from the impacted a
rea
tukey = pairwise_tukeyhsd(endog=df2['PFOA'],      # Data
                          groups=df2['Source'],   # Groups
                          alpha=0.05)          # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[126]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Apple_Hill-Impacted	Barr-Impacted	-5.0889	0.001	-8.426	-1.7517	True
Apple_Hill-Impacted	ThisStudy-Impacted	-1.6038	0.5962	-4.9989	1.7913	False
Apple_Hill-Impacted	VT-DEC-Impacted	-3.1692	0.0706	-6.5168	0.1784	False
Barr-Impacted	ThisStudy-Impacted	3.4851	0.0015	1.0657	5.9045	True
Barr-Impacted	VT-DEC-Impacted	1.9196	0.1508	-0.4326	4.2719	False
ThisStudy-Impacted	VT-DEC-Impacted	-1.5655	0.3414	-3.9993	0.8683	False



SI Figure 5: Soil PFOA Tukey variance overlap between studies in impacted area only The result of this smaller Tukey analysis is the same as the one that included all six groups, though there is now slightly more overlap in the variance between the VT-DEC samples and those of the Barr study.

Below, we run ANOVA and Tukey analyses on soil PFOS concentration with all six sample groups

```
In [127]: #run one-way ANOVA test on the sample groups (without the peripheral are
          a groups)
          lm = ols('PFOS ~ Source',data=df).fit()
          table = sm.stats.anova_lm(lm)
          print(table)
```

```
             df      sum_sq   mean_sq         F    PR(>F)
Source      5.0   21.578332  4.315666  5.694771  0.000059
Residual  218.0  165.206881  0.757830       NaN       NaN
```

The p value of 0.000059 is below 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

In [128]:

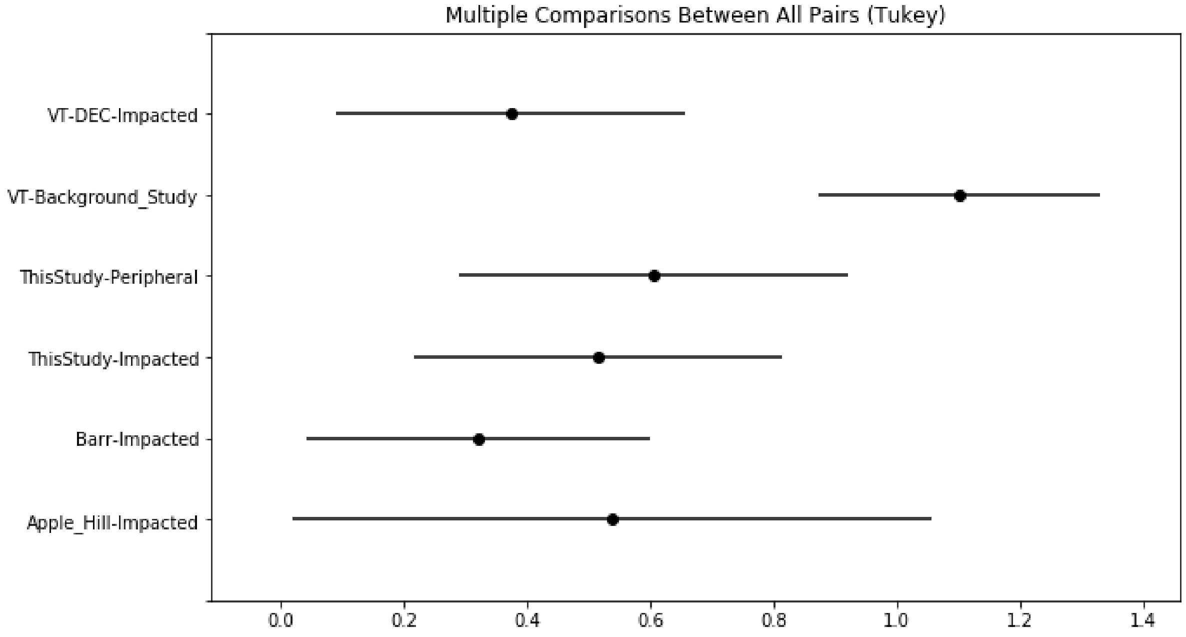
```
#Run the Tukey analysis on all six sample groupings for PFOS
tukey = pairwise_tukeyhsd(endog=df['PFOS'],      # Data
                          groups=df['Source'],   # Groups
                          alpha=0.05)           # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[128]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Apple_Hill-Impacted	Barr-Impacted	-0.2165	0.9	-1.0156	0.5826	False
Apple_Hill-Impacted	ThisStudy-Impacted	-0.0227	0.9	-0.8357	0.7902	False
Apple_Hill-Impacted	ThisStudy-Peripheral	0.0673	0.9	-0.7597	0.8944	False
Apple_Hill-Impacted	VT-Background_Study	0.5632	0.2748	-0.1963	1.3227	False
Apple_Hill-Impacted	VT-DEC-Impacted	-0.1644	0.9	-0.966	0.6372	False
Barr-Impacted	ThisStudy-Impacted	0.1937	0.9	-0.3856	0.773	False
Barr-Impacted	ThisStudy-Peripheral	0.2838	0.7221	-0.3151	0.8827	False
Barr-Impacted	VT-Background_Study	0.7797	0.001	0.2781	1.2812	True
Barr-Impacted	VT-DEC-Impacted	0.0521	0.9	-0.5112	0.6154	False
ThisStudy-Impacted	ThisStudy-Peripheral	0.0901	0.9	-0.5272	0.7074	False
ThisStudy-Impacted	VT-Background_Study	0.586	0.0183	0.0626	1.1093	True
ThisStudy-Impacted	VT-DEC-Impacted	-0.1416	0.9	-0.7244	0.4412	False
ThisStudy-Peripheral	VT-Background_Study	0.4959	0.0977	-0.0491	1.0409	False
ThisStudy-Peripheral	VT-DEC-Impacted	-0.2317	0.8691	-0.834	0.3706	False
VT-Background_Study	VT-DEC-Impacted	-0.7276	0.001	-1.2331	-0.222	True



SI Figure 6: Soil PFOS Tukey variance overlap between studies This analysis shows that soil PFOS concentration in the Vermont Background Study is significantly higher than that of several of the other studies performed in the Bennington area, and that of our samples collected from peripheral regions.

For completeness, we will rerun the ANOVA analysis on just the four impacted sample groups in the Bennington area.

```
In [129]:
#run one-way ANOVA test on the sample groups (without the peripheral area groups)
lm = ols('PFOS ~ Source',data=df2).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
             df     sum_sq   mean_sq         F    PR(>F)
Source      3.0   0.964096  0.321365  1.168441  0.324638
Residual  123.0  33.829610  0.275037       NaN       NaN
```

The p value of 0.325 is above 0.05, so we accept the null hypothesis that all groups are similar.

These two analyses indicate statistical similarity in soil PFOS concentration between all sample groups, except that of the Vermont Background study conducted by UVM on behalf of VT-DEC. This study contains several samples with high PFOS. With regard to the data used in this study, the analysis suggests consistency in sample concentrations between all studies.

Soil PFOA Retention and Land Cover

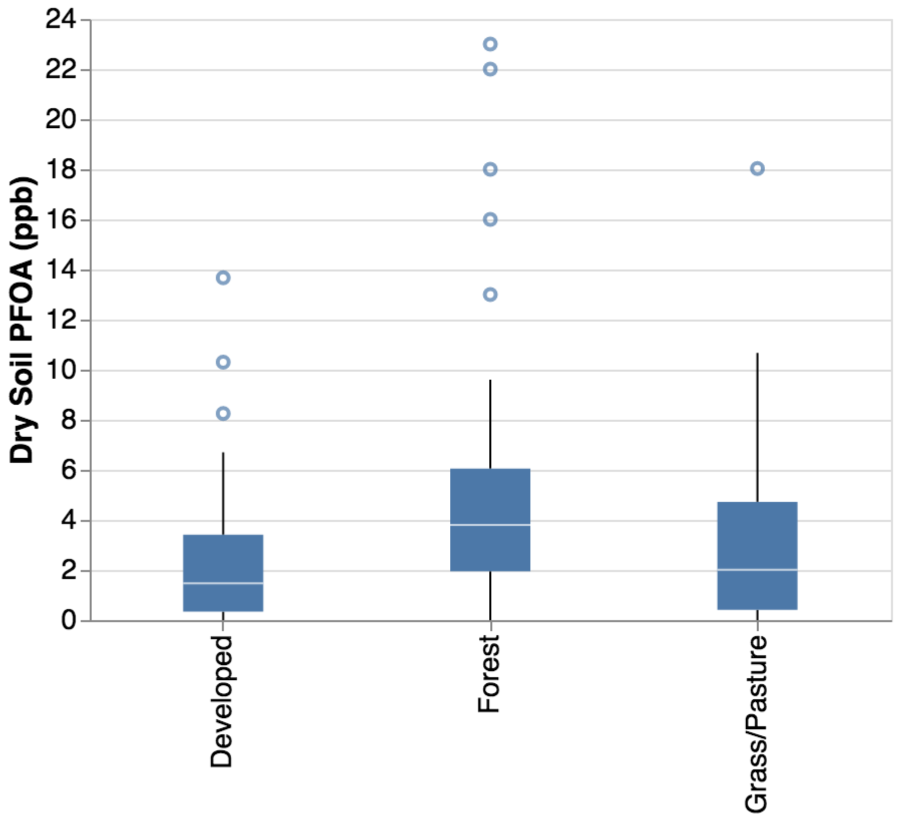
Below, we explore how land cover may be related to the degree of PFAS retention in soils in the Bennington area. We use this study's data and data from the multiple studies discussed above from the impacted region around Bennington.

land cover was determined by using GIS to intersect the soil sample points with the 2016 USGS National Land Cover Dataset, and lumping land cover into three categories, Forest, Developed/Barren, and Grassland/Pasture.

In [130]:

```
#make simple boxplot of three groups of land use/cover
ytitle=" Dry Soil PFOA (ppb)"
xtitle=""
alt.Chart(df2).mark_boxplot(size=40).encode(
    x=alt.X('LandCover', title=xtitle),
    y=alt.Y('PFOA', title=ytitle)).properties(width=400, height=300).con
figure_axis(labelFontSize=14,
    titleFontSize=15)
```

Out[130]:



SI Figure 7: Boxplots of Soil PFOA Concentration in Bennington Impacted area This includes data from this study and the other three studies from the Bennington area referenced above.

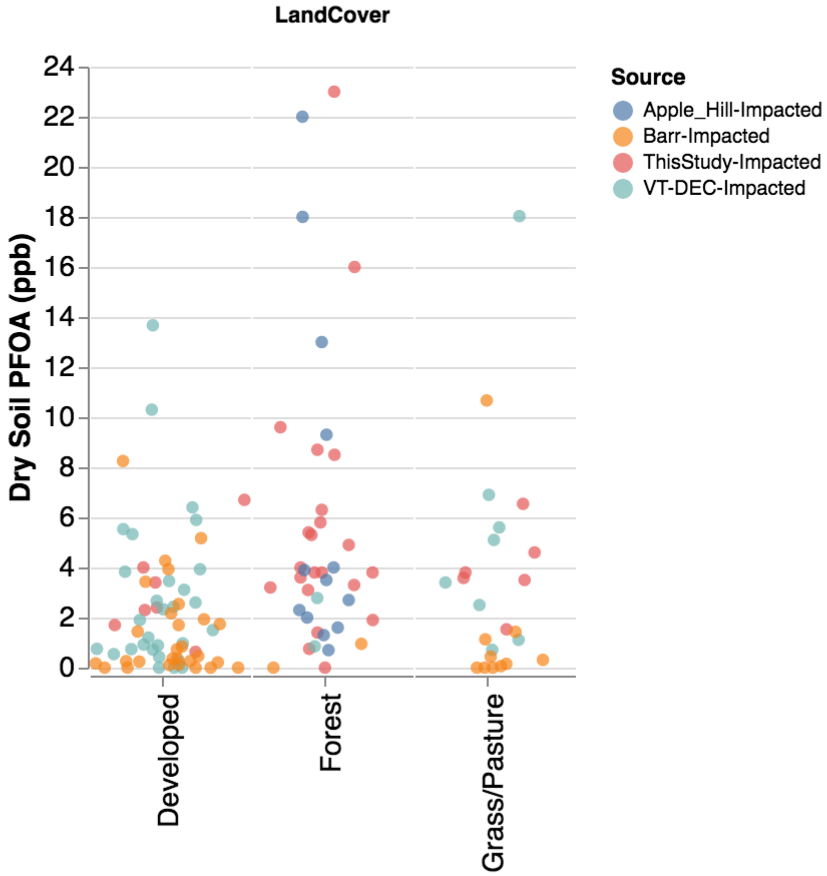
To help visualize the data more completely, we inlude a "strip plot" graph that shows each individual data point color coded by the study from which it was derived.

In [131]:

```
#create the stripplot color coded by study origin
stripplot =  alt.Chart(df2, width=80).mark_circle(size=40).encode(
    x=alt.X(
        'jitter:Q',
        title=None,
        axis=alt.Axis(values=[0], ticks=True, grid=False, labels=False),
        scale=alt.Scale(),
    ),
    y=alt.Y('PFOA', title = 'Dry Soil PFOA (ppb)'),
    color=alt.Color('Source'),
    column=alt.Column(
        'LandCover',
        header=alt.Header(
            labelAngle=-90,
            titleOrient='top',
            labelOrient='bottom',
            labelAlign='right',
            labelPadding=3,
            labelFontSize=14,
        ),
    ),
).transform_calculate(
    # Generate Gaussian jitter with a Box-Muller transform
    jitter='sqrt(-2*log(random()))*cos(2*PI*random())'
).configure_facet(
    spacing=0
).configure_view(
    stroke=None
).configure_axis(labelFontSize=14,
    titleFontSize=14)

stripplot
```

Out[131]:



SI Figure 8: Strip Plot of Soil PFOA Concentration in Bennington Impacted area This includes data from this study and the other three studies from the Bennington area referenced above.

We run ANOVA and Tuckey analyses to test for statistictal difference in soil PFOA concentration in areas of different land cover.

In [132]:

```
#run one-way ANOVA test on the sample groups (without the peripheral are
a groups)
lm = ols('PFOA ~ LandCover',data=df2).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
              df       sum_sq     mean_sq         F    PR(>F)
LandCover    2.0   257.988790  128.994395  7.676051  0.00072
Residual   124.0  2083.793454   16.804786       NaN       NaN
```

The p value of 0.00072 is below 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

In [133]:

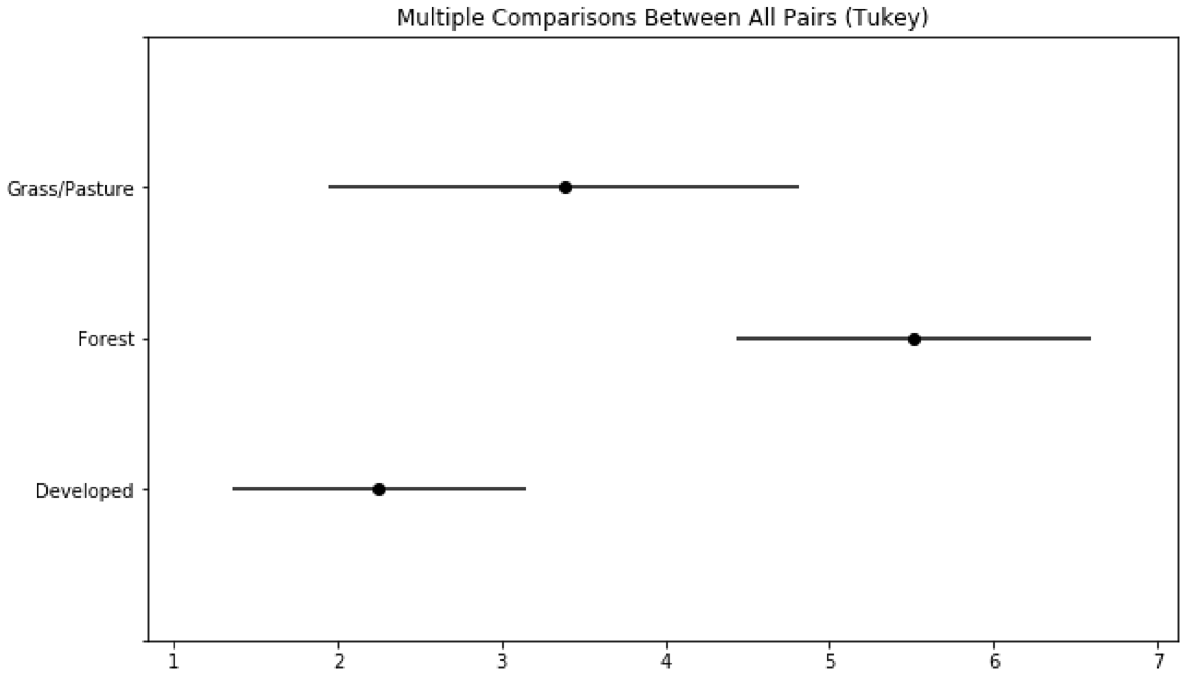
```
#Run the Tukey analysis on land cover and PFOA in the impacted zone
tukey = pairwise_tukeyhsd(endog=df2['PFOA'],      # Data
                          groups=df2['LandCover'],    # Groups
                          alpha=0.05)          # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[133]:

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05

    group1        group2  meandiff   p-adj    lower   upper  reject
 Developed        Forest    3.2614   0.001   1.2858   5.237    True
 Developed Grass/Pasture    1.1266  0.4879  -1.2012  3.4544   False
    Forest Grass/Pasture   -2.1348  0.1148  -4.6579  0.3882   False
```



SI Figure 9: Soil PFOA Tukey variance overlap between sample point land cover

This analysis indicates that PFOA soil levels in developed or barren land areas are significantly lower than that of forested areas, but not significantly lower than grassland areas. This may be due to higher PFOA retention in the forest soils with higher organic carbon content, or possibly more scavenging of PFOA from the air in by tree canopy.

Because the Barr study data included a large number of sample sites on developed land cover, this could skew the above analysis. It therefore seems more appropriate to re-run the analysis without the Barr data, and include just the other three studies.

In [134]:
```
#remove the Barr data from the dataframe
df3 = df2[df2['Source'] != 'Barr-Impacted']
```

In [135]:
```
#run one-way ANOVA test on the sample groups (without the peripheral are a groups, and without Barr data)
lm = ols('PFOA ~ LandCover',data=df3).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
             df        sum_sq    mean_sq         F    PR(>F)
LandCover   2.0    156.922267  78.461134  3.812151  0.026018
Residual   84.0   1728.875694  20.581853       NaN       NaN
```

The p value of 0.02 is below 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

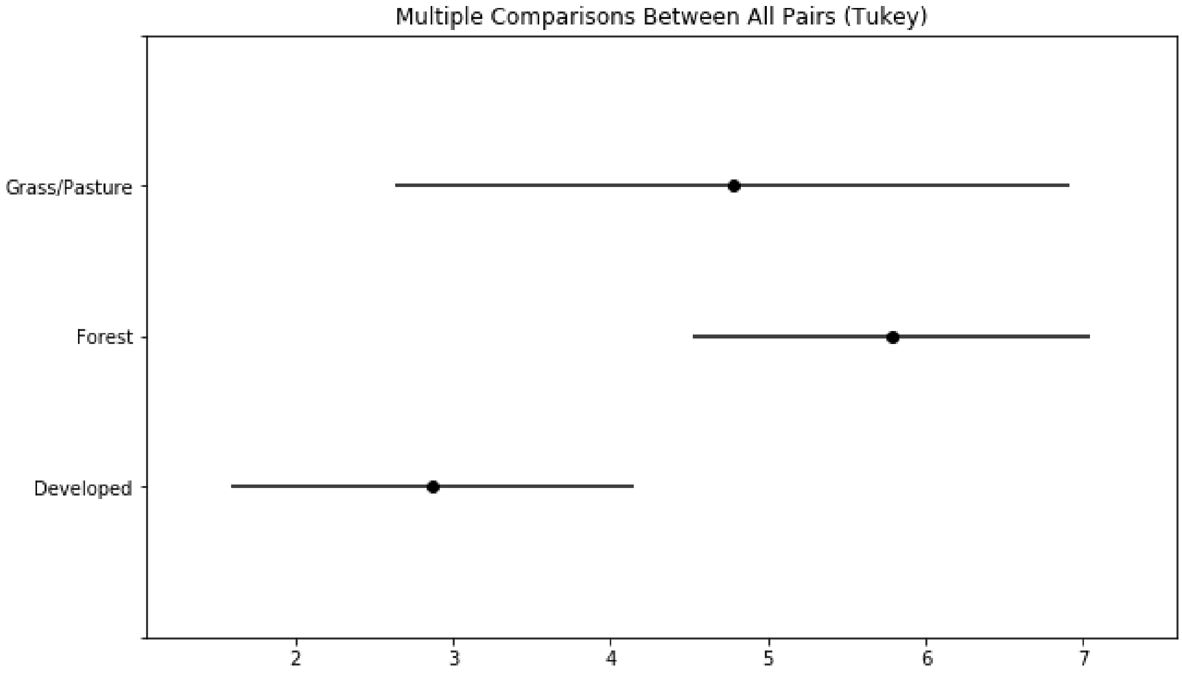
In [136]:
```
tukey = pairwise_tukeyhsd(endog=df3['PFOA'],      # Data
                          groups=df3['LandCover'],   # Groups
                          alpha=0.05)            # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[136]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Developed	Forest	2.9123	0.0202	0.3782	5.4465	True
Developed	Grass/Pasture	1.905	0.3823	-1.5045	5.3144	False
Forest	Grass/Pasture	-1.0074	0.7399	-4.4039	2.3892	False



SI Figure 10: Soil PFOA Tukey variance overlap between sample point land cover - Without data from Barr Study

Without the Barr data included in the analysis, the developed/barren land has significantly lower soil PFOA concentration than in forested areas. Because the Barr study included many more samples sites in developed areas, this relationship may help explain why soil PFOA levels from the Barr study are significantly lower than those of this study's data and those of the other two Bennington area sample sets.

Analysis of Difference Between Soil Sampling Regions in This Study

In order to test the hypothesis that industrial air emission of PFOA in the Bennington/Hoosick Falls area impacted soil in the Benington Local and Downwind sampling areas relative to other, we plot and perform statistical analysis of the soil PFOA and PFOS concentrations between the different sampling regions.

In [140]:

```
#read the datafile
regdf = pd.read_csv("All_Benn_Data_4.csv")
regdf.head(10)
```

Out[140]:

	Region	Easting	Northing	PFOA	PFOS	PFHpA	PFHxA	PFNA	Sum_PFAS
0	Downwind	651666	4753625	5.3	1.40	0.45	0.00	0.38	8.36
1	Downwind	652168	4753146	3.2	0.32	0.19	0.00	0.00	3.71
2	Downwind	652213	4753238	3.1	0.00	0.21	0.00	0.00	3.31
3	Downwind	652828	4752540	3.6	0.50	0.37	0.20	0.00	4.67
4	Downwind	652940	4752676	6.3	0.48	0.39	0.23	0.00	7.40
5	Downwind	649567	4750710	3.8	0.46	0.39	0.43	0.20	5.28
6	Downwind	649655	4751071	1.4	0.41	0.00	0.00	0.00	1.81
7	Downwind	649844	4751752	8.7	0.86	0.49	0.32	0.00	10.37
8	Downwind	650039	4751757	4.0	0.62	0.24	0.20	0.00	5.06
9	Downwind	649307	4750384	3.3	0.00	0.27	0.26	0.16	3.99

In [141]:

```
#examine the data, find the uniqe regions
regdf['Region'].unique()
```

Out[141]:

```
array(['Downwind', 'North', 'Upwind', 'Far', 'Local ', 'Taconic'],
      dtype=object)
```

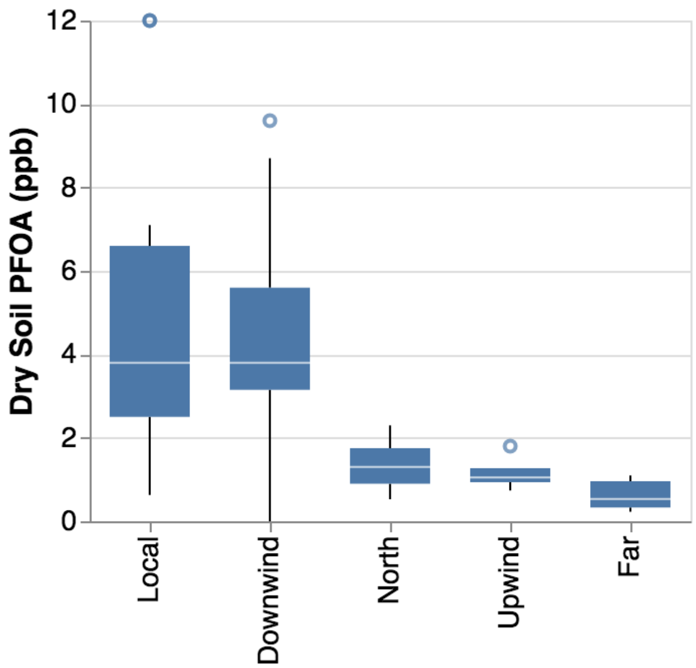
In [143]:

```
#remove the Taconic region, which is not analyzed in the context of this
work
regdf = regdf[regdf['Region'] !='Taconic']
```

In [144]:

```
#Create boxplot of soil PFOA concentrations
ytitle=" Dry Soil PFOA (ppb)"
xtitle=""
alt.Chart(regdf).mark_boxplot(size=40).encode(
    x=alt.X('Region', title=xtitle, sort=['Local ','Downwind','North','Upwind','Far']),
    y=alt.Y('PFOA', title=ytitle, scale=alt.Scale(
            domain=(0, 12),
            clamp=True))).properties(width=300, height=250).configure_axis(labelFontSize=14,
    titleFontSize=15)
```

Out[144]:



SI Figure 11: Boxplots of soil PFOA concentrations across sampling regions

Boxplots show higher PFOA concentration in the Bennington Local & Downwind regions relative to other regions. We will run the ANOVA test for statistical difference.

In [145]:

```
#run one-way ANOVA test on the sample regions soil PFOA concentration
lm = ols('PFOA ~ Region',data=regdf).fit()
table = sm.stats.anova_lm(lm)
print(table)
```

```
            df      sum_sq    mean_sq         F    PR(>F)
Region     4.0  239.312087  59.828022  7.837711  0.000038
Residual  60.0  458.001211   7.633354       NaN       NaN
```

The p value of 0.000038 is below 0.05, so we reject the hypothesis that all groups are similar, and proceed to the Tukey pairwise test.

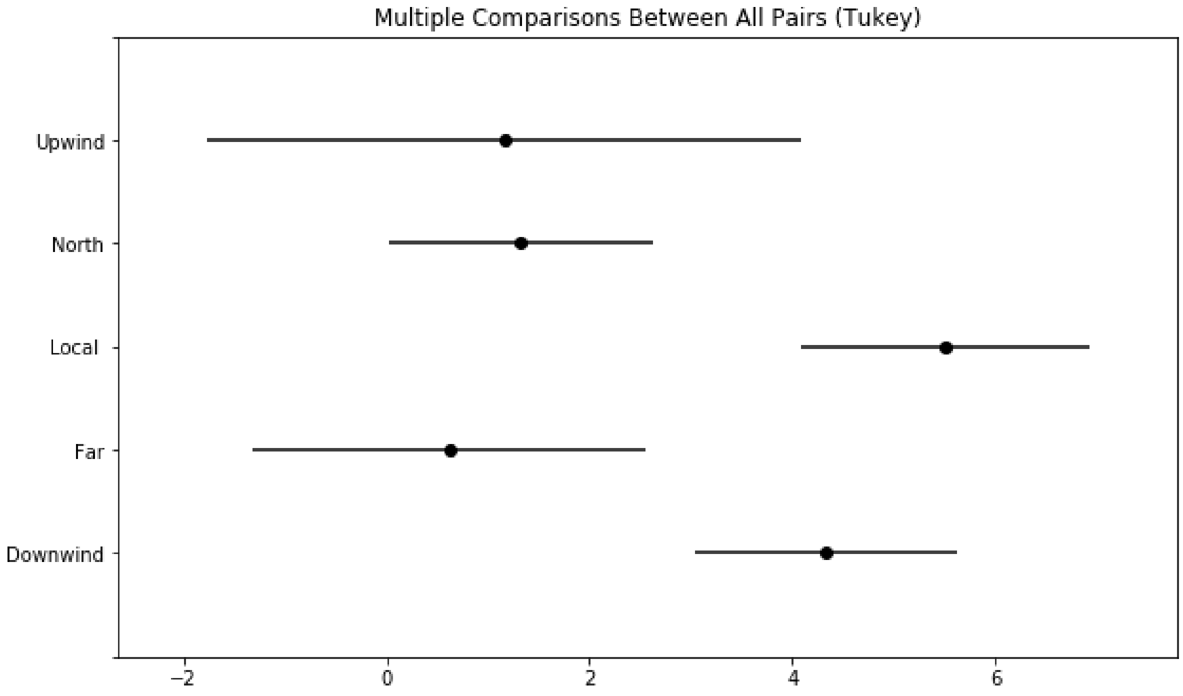
In [147]:
```
tukey = pairwise_tukeyhsd(endog=regdf['PFOA'],      # Data
                          groups=regdf['Region'],   # Groups
                          alpha=0.05)             # Significance level

tukey.plot_simultaneous()    # Plot group confidence intervals
plt.vlines(x=49.57,ymin=-0.5,ymax=4.5, color="red")
tukey.summary()
```

Out[147]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Downwind	Far	-3.7114	0.0187	-6.9864	-0.4365	True
Downwind	Local	1.1731	0.7093	-1.5108	3.8569	False
Downwind	North	-3.0026	0.0118	-5.5237	-0.4816	True
Downwind	Upwind	-3.1689	0.2399	-7.4436	1.1057	False
Far	Local	4.8845	0.0014	1.4826	8.2864	True
Far	North	0.7088	0.9	-2.5661	3.9837	False
Far	Upwind	0.5425	0.9	-4.2159	5.3009	False
Local	North	-4.1757	0.001	-6.8595	-1.4918	True
Local	Upwind	-4.342	0.0525	-8.7146	0.0306	False
North	Upwind	-0.1663	0.9	-4.441	4.1083	False

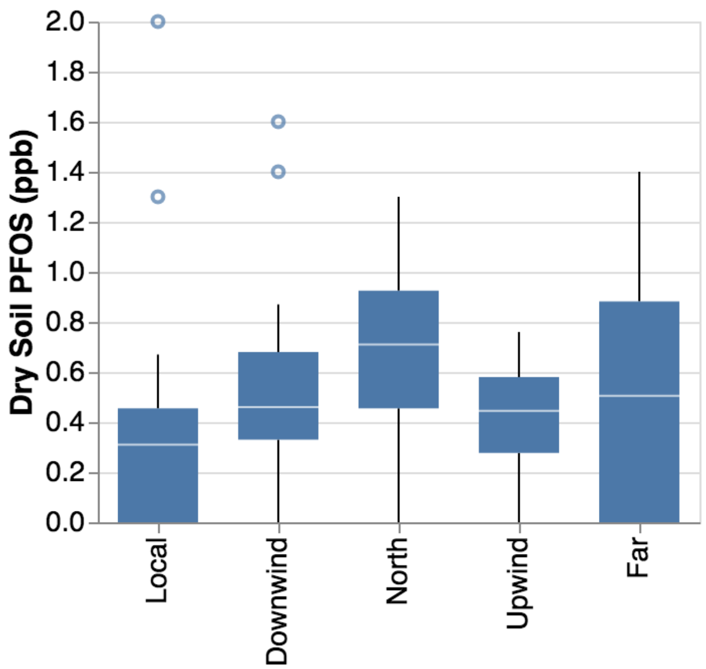


SI Figure 12: Tukey analysis of soil PFOA concentration between sampling regions This analysis shows that soil PFOA concentration in the Local and Downwind regions is significantly higher than that of the North of Wind Pattern and Far-Affield regions. There is statistical overlap between the Upwind region and all other sample regions. This is likely due in part to the small number of samples (four) there causing higher variance.

Run analyses for soil PFOS concentrations

```
In [148]: #Create boxplot of soil PFOS concentrations
          ytitle=" Dry Soil PFOS (ppb)"
          xtitle=""
          alt.Chart(regdf).mark_boxplot(size=40).encode(
              x=alt.X('Region', title=xtitle, sort=['Local ','Downwind','North','U
          pwind','Far']),
              y=alt.Y('PFOS', title=ytitle, scale=alt.Scale(
                      domain=(0, 2),
                      clamp=True))).properties(width=300, height=250).configure_ax
          is(labelFontSize=14,
              titleFontSize=15)
```

Out[148]:



SI Figure 12: Boxplots of soil PFOS concentrations across sampling regions

There is not an apparent difference in PFOA concentration between the sampling regions. We will run the ANOVA test regardless.

```
In [149]: #run one-way ANOVA test on the sample regions soil PFOS concentration
          lm = ols('PFOS ~ Region',data=regdf).fit()
          table = sm.stats.anova_lm(lm)
          print(table)
```

```
            df      sum_sq   mean_sq         F    PR(>F)
Region     4.0    3.259154  0.814789  0.354856  0.839657
Residual  60.0  137.766560  2.296109       NaN       NaN
```

The p value of 0.8397 is greater than 0.05, so we accept the hypothesis that all groups are similar. There is no significant difference in soil PFOS concentration between any sample regions.