Sample Python Data Projects

SEO, Machine Learning, and Data Analysis

Internal Linking

Check every page for internal linking opportunities based on keyword.

```python
dfkeywords = pd.read_csv('/Users/jessefortson/Desktop/keywords_list.csv')
dfurls = pd.read_csv('/Users/jessefortson/Desktop/blog_url.csv')
dfurls.head()
```

Python

	URL
0	https://www.trainerroad.com/blog/interbike-2011/
1	https://www.trainerroad.com/blog/version-0-5-1/
2	https://www.trainerroad.com/blog/an-update-on-...
3	https://www.trainerroad.com/blog/cycleops-supe...
4	https://www.trainerroad.com/blog/beta-ending-s...

```python
#Make list of keywords
keywordsandpages = dfkeywords[["Keyword","URL"]].values.tolist()
keywordspages=[]
firstkeywords=[]
for keywordandpage in keywordsandpages:
    if(keywordandpage[0] not in firstkeywords):
        keywordspages.append([keywordandpage[0],keywordandpage[1]])
        firstkeywords.append(keywordandpage[0])
```

Python

```python
#Make list of URLS
addresses = dfurls["URL"].values.tolist()
urls=[]
for address in addresses:
    if("?" not in address):
        urls.append(address)
urls
```

Python

Output exceeds the size limit. Open the full output data in a text editor

```
['https://www.trainerroad.com/blog/interbike-2011/',
 'https://www.trainerroad.com/blog/version-0-5-1/',
 'https://www.trainerroad.com/blog/an-update-on-virtualpower-for-the-cycleops-fluid-2/',
 ...ycleops-supermagneto-pro-added/',
 ...eta-ending-soon/',
 ...hree-new-workouts-2x20-sst-and-endurance/',
 ...sk-chad-volume-in-training-plans/',
 ...sk-chad-power-vs-heartrate/',
 ...-quit-my-job/',
 ...sk-chad-racing-in-early-base/',
 ...sk-chad-cadence-during-workouts-and-rest-intervals/',
 ...lueberry-protein-smoothie/',
 ...nitial-outdoor-workout-testing/',
 ...improved-powerbeam-pro-support/',
 'https://www.trainerroad.com/blog/endurance-films-on-trainerroad/',
 'https://www.trainerroad.com/blog/workout-page-update/',
 'https://www.trainerroad.com/blog/workout-of-the-week-kettle/',
 'https://www.trainerroad.com/blog/scott-waters-the-man-who-bikes/',
 'https://www.trainerroad.com/blog/293/',
 'https://www.trainerroad.com/blog/workout-of-the-week-disaster/',
 'https://www.trainerroad.com/blog/advice-for-beginning-cyclists/',
 'https://www.trainerroad.com/blog/which-assessment-protocol-should-i-use/',
 'https://www.trainerroad.com/blog/im-training-on-trainerroad/',
 'https://www.trainerroad.com/blog/aero-position-assessment/',
 'https://www.trainerroad.com/blog/recovery-weeks/',
...
```

	sourceurl	Search query	Landing page	Number links
0	https://www.trainerroad.com/blog/an-update-on-...	power curve	https://www.trainerroad.com/blog/how-to-use-yo...	0
1	https://www.trainerroad.com/blog/three-new-wor...	sweet spot training	https://www.trainerroad.com/blog/sweet-spot-tr...	0
2	https://www.trainerroad.com/blog/three-new-wor...	sweet spot intervals	https://www.trainerroad.com/blog/sweet-spot-tr...	0
3	https://www.trainerroad.com/blog/three-new-wor...	spot training	https://www.trainerroad.com/blog/sweet-spot-tr...	0
4	https://www.trainerroad.com/blog/ask-chad-cade...	threshold power	https://www.trainerroad.com/blog/what-ftp-real...	0

Keyword Cannibalization

Use the Google Search Console API to identify keyword cannibalisation across pages.

query	unique_pages	total_clicks	total_impressions	avg_ctr	avg_position
ftp cycling	4	162	8358	9.965000	4.545000
how to increase ftp	4	35	357	38.827500	2.460000
ftp test	4	25	4272	4.785000	6.830000
when to do an ftp test	3	7	31	46.296667	2.223333
what does ftp stand for in cycling	3	5	323	11.726667	6.463333
ftp vs vo2 max	3	34	96	49.476667	1.103333
what is ftp cycling	3	32	3543	33.953333	6.783333
trainer road strength training	3	6	21	28.573333	2.810000
sweet spot ftp	3	18	254	20.990000	3.826667
how to improve ftp	3	33	271	43.900000	2.820000

```
df[df['query']=='ftp test'].sort_values(by='impressions', ascending=False).head(20)
```

	query	page	clicks	impressions	ctr	position
945	ftp test	https://www.trainerroad.com/blog/ftp-assessment-tips/	14	3059	0.46	3.93
2667	ftp test	https://www.trainerroad.com/blog/what-ftp-really-means-to-cyclists/	6	1121	0.54	11.06
12789	ftp test	https://www.trainerroad.com/blog/new-ramp-test-makes-ftp-testing-more-efficient-and-less-stressful/	1	68	1.47	10.25
4056	ftp test	https://www.trainerroad.com/blog/is-my-ftp-too-low/	4	24	16.67	2.08

```
def query(service, site_url, payload):
    """Run a query on the Google Search Console API and return a dataframe of results.

    Args:
        service (object): Service object from connect()
        site_url (string): URL of Google Search Console property
        payload (dict): API query payload dictionary

    Return:
        df (dataframe): Pandas dataframe containing requested data.

    """

    response = service.searchanalytics().query(siteUrl=site_url, body=payload).execute()

    results = []

    for row in response['rows']:
        data = {}

        for i in range(len(payload['dimensions'])):
            data[payload['dimensions'][i]] = row['keys'][i]

        data['clicks'] = row['clicks']
        data['impressions'] = row['impressions']
        data['ctr'] = round(row['ctr'] * 100, 2)
        data['position'] = round(row['position'], 2)
        results.append(data)

    return pd.DataFrame.from_dict(results)
```

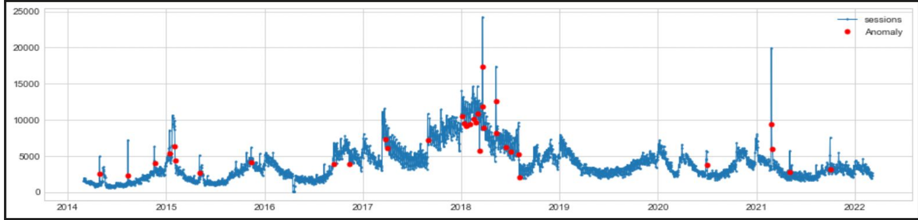
```
payload = {
    'startDate': "2022-02-01",
    'endDate': "2022-03-01",
    'dimensions': ["query", "page", "date"],
    'rowLimit': 20000,
    'startRow': 0
}

site_url = [redacted]

df = query(service, site_url, payload)
df.head()
```

	query	page	date	clicks	impressions	ctr	position
0	trainerroad	https://www.trainerroad.com/	2022-02-09	644	1084	59.41	1.0
1	trainerroad	https://www.trainerroad.com/	2022-02-10	628	1061	59.19	1.0
2	trainerroad	https://www.trainerroad.com/	2022-02-08	608	1007	60.38	1.0
3	trainerroad	https://www.trainerroad.com/	2022-02-01	598	1028	58.17	1.0
4	trainerroad	https://www.trainerroad.com/	2022-02-15	591	988	59.82	1.0

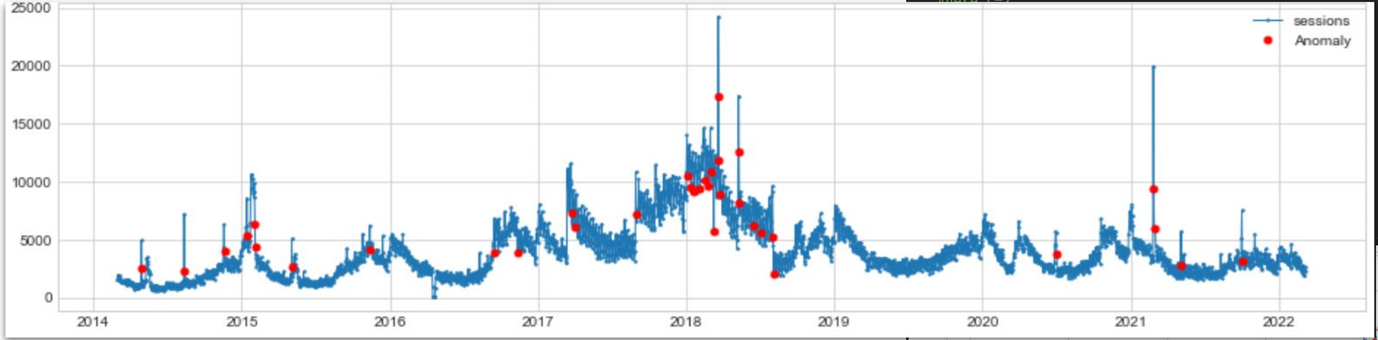
Google Analytics Anomaly Detection

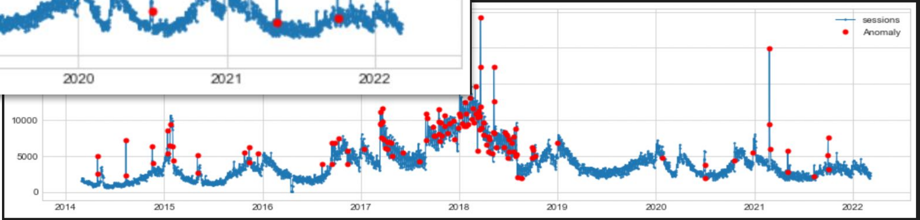
Using the Google Analytics API to to detect a wide range of anomaly types.

```
#Persist anomaly detection
#Uses a double rolling aggregate model and find anomolaies from previous time series.
#Change c=3.0 to increase the previous TS values, lowering increases sensativity

persist_ad = PersistAD(c=3, side='negative')
anomalies = persist_ad.fit_detect(s)

chart = plot(s,
     anomaly=anomalies,
     ts_linewidth=1,
     ts_markersize=3,
     anomaly_markersize=5,
     anomaly_color='red',
     anomaly_tag='marker')
```




```
#Same as above, but side='both' looks for anomalies in both direction, change be changed to positive or negative
#here c=2
```





Forecasting Sales with Machine Learning