Consumer & Visitor Insights For Neighborhoods

May 4, 2020

```
In [1]: import pandas as pd
        import numpy as np
```

```
In [2]: path = "./dataset/Consumer Insights/"
        data_df = pd.read_csv(path+'Consumer Insights.csv', index_col=0)
```

0.1

```
In [3]: data_df.head()
```

Out[3]:

census_block_group	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	visitor_home_cbgs	visitor_work_cbgs
1.005951e+10	1538352000	1541030400	75122.0	18314.0	{"010059501003":127,"010059509001":111,"010059...	{"010059501003":109,"010810407002":62,"0108104...
1.009051e+10	1538352000	1541030400	95649.0	38942.0	{"010730113021":210,"010090506022":205,"010090...	{"010890111001":271,"010730045001":269,"010439...
1.047957e+10	1538352000	1541030400	14009.0	3039.0	{"010479567011":67,"010479567021":60}	
1.069040e+10	1538352000	1541030400	128169.0	25418.0	{"010690402013":370,"010690402011":322,"010690...	
1.073011e+10	1538352000	1541030400	51453.0	9499.0	{"010090507001":183,"010730113021":167,"010730...	

```
1.047957e+10                                    {"010479567021":52}
1.069040e+10        {"010690402024":313,"010690415004":203,"010450...
1.073011e+10        {"010730045001":140,"010730027001":123,"010730...

                   distance_from_home  \
census_block_group
1.005951e+10                194724.0
1.009051e+10                120587.0
1.047957e+10                 67774.0
1.069040e+10                 42684.0
1.073011e+10                 18878.0

                                          related_same_day_brand  \
census_block_group
1.005951e+10        ["Chick-fil-A","mcdonalds","Marathon Petroleum...
1.009051e+10        ["Shell Oil","mcdonalds","Chick-fil-A","Chevron"]
1.047957e+10                                    ["Dollar General"]
1.069040e+10        ["Chick-fil-A","Sam's Club","Dollar General","...
1.073011e+10                  ["Chevron","Daylight Donuts","walmart"]

                                         related_same_month_brand  \
census_block_group
1.005951e+10        ["walmart","mcdonalds","Dollar General","Chick...
1.009051e+10        ["walmart","mcdonalds","Shell Oil","Chick-fil-...
1.047957e+10        ["walmart","Dollar General","mcdonalds","Chevr...
1.069040e+10        ["walmart","Dollar General","mcdonalds","Marat...
1.073011e+10        ["walmart","Chevron","Dollar General","Shell O...

                                                       top_brands  \
census_block_group
1.005951e+10             ["CrossFit","Health Mart","Coldwell Banker"]
1.009051e+10                                                      []
1.047957e+10                                    ["Dollar General"]
1.069040e+10        ["Chick-fil-A","Sam's Club","Olive Garden","mc...
1.073011e+10                                  ["Chevron","CrossFit"]

                                                   popularity_by_hour  \
census_block_group
1.005951e+10        [2617,2457,2403,2519,2646,3007,3886,7566,5508,...
1.009051e+10        [6556,6325,6222,6355,6586,7350,8568,8099,7378,...
1.047957e+10        [807,790,796,786,851,951,1134,1797,1355,1241,1...
1.069040e+10        [2121,1828,1784,1704,1861,2373,3730,7497,7093,...
1.073011e+10        [3804,3716,3686,3672,3735,4115,4855,5946,4526,...

                                                    popularity_by_day
census_block_group
1.005951e+10        {"Monday":12000,"Tuesday":12224,"Wednesday":10...
1.009051e+10        {"Monday":12125,"Tuesday":12984,"Wednesday":12...
```

```
        1.047957e+10        {"Monday":2314,"Tuesday":2340,"Wednesday":2195...
        1.069040e+10        {"Monday":21141,"Tuesday":21143,"Wednesday":17...
        1.073011e+10        {"Monday":8402,"Tuesday":8414,"Wednesday":8550...
```

```
In [6]: data_df.dtypes
```

```
Out[6]: date_range_start              int64
        date_range_end                int64
        raw_visit_count             float64
        raw_visitor_count           float64
        visitor_home_cbgs            object
        visitor_work_cbgs            object
        distance_from_home          float64
        related_same_day_brand       object
        related_same_month_brand     object
        top_brands                   object
        popularity_by_hour           object
        popularity_by_day            object
        dtype: object
```

```
In [7]: data_df.shape
```

```
Out[7]: (220735, 12)
```

```
In [10]: data_df['related_same_day_brand'].value_counts()
```

```
Out[10]: []
         ["Dunkin' Donuts"]
         ["starbucks"]
         ["walmart"]
         ["mcdonalds"]
         ["Wawa"]
         ["7-Eleven US"]
         ["Publix Super Markets"]
         ["QuikTrip"]
         ["Kroger"]
         ["Chick-fil-A"]
         ["Chevron"]
         ["Sheetz"]
         ["Safeway"]
         ["Costco Wholesale Corp."]
         ["Meijer"]
         ["ARCO"]
         ["H-E-B"]
         ["Kwik Trip"]
         ["Circle K Stores"]
         ["King Soopers"]
         ["Speedway"]
         ["Circle K Stores","QuikTrip"]
```

```
["Shell Oil"]
["Dunkin' Donuts","Stop & Shop"]
["Giant Eagle"]
["Wawa","Dunkin' Donuts"]
["Family Dollar Stores"]
["Target"]
["mcdonalds","Speedway"]

["starbucks","CVS","Shake Shack"]
["Hy-Vee","mcdonalds","Casey's General Stores","Love's Travel Stops and Country Stores
["Baja Fresh","starbucks","Auto Wares","mcdonalds","Kwik Trip","Freshťns"]
["Dunkin' Donuts","Cinemark Theatres","Sunoco","Dollar Tree"]
["Kwik Shop","Dillons Supermarkets","walmart","Sonic","mcdonalds"]
["walmart","Chick-fil-A","Baymont Inn & Suites","Jack's Family Restaurants","Dollar Ge
["Sheetz","Giant Eagle","Dollar Tree","Eat'n Park"]
["SHOP Äôn SAVE Supermarkets","walmart","Dollar General","Sheetz","mcdonalds","Kroger"
["Family Dollar Stores","Chevron","Health Mart","Sinclair Oil","walmart","Maverik","Ex
["walmart","Samurai Sam's Teriyaki Grill","Safeway","Chevron","Habitat for Humanity","
["Family Dollar Stores","Boys & Girls Clubs of America","walmart","Valero Energy"]
["SonnyÄôs Real Pit Bar-B-Q","walmart","Family Dollar Stores","Save-A-Lot (Onex)","Piz
["Target","Plato's Closet"]
["Hobby Lobby Stores","walmart","Bealls Tx","Sonic","United Supermarkets","Pizza Hut"
["Chick-fil-A","mcdonalds","walmart","Jack's Family Restaurants","Dollar General"]
["starbucks","Fred Meyer Jewelers","WinCo Foods","Dollar Tree","Fred Meyer","Dutch Bro
["Zumiez","Costco Wholesale Corp.","Petco"]
["mcdonalds","Marathon Petroleum","Owen's Market","BP"]
["Dairy Queen","Casey's General Stores","mcdonalds"]
["walmart","United Supermarkets","Taco Casa","Whataburger"]
["Shell Oil","Health Mart","Dollar General","mcdonalds"]
["Kwik Trip","Casey's General Stores","SuperAmerica","Holiday Station"]
["Casey's General Stores","walmart","SUBWAY","mcdonalds","Dollar General"]
["Sheetz","Goodwill Industries","mcdonalds","walmart","Ford Motor Company"]
["walmart","Sonic","Marathon Petroleum","Shell Oil","Dollar General","Hardee's","Exxon
["76","Safeway","Chevron","A&W Restaurants","Taco Bell"]
["Marathon Petroleum","Family Dollar Stores","walmart","mcdonalds","United States Post
["Dollar Tree","Dunkin' Donuts","mcdonalds","walmart"]
["mcdonalds","Wawa","Fred Meyer Jewelers","7-Eleven US","Chick-fil-A","Kroger","Navy
["walmart","Big Lots Stores","Chevron","Jack's Family Restaurants","Shell Oil","Chick-
Name: related_same_day_brand, Length: 73198, dtype: int64
```

In [11]:
```
data_df['related_same_month_brand'].value_counts()
```

Out[11]:
```
[]
["SmartStyle Family Hair Salons"]
["Department of Veterans Affairs"]
["H&R Block"]
["Sprint"]
["MasterCuts"]
```

```
["Supercuts"]
["Dunkin' Donuts"]
["Safeway"]
["Yogen Fržz"]
["Yogen Fržz","Department of Veterans Affairs"]
["CVS","SmartStyle Family Hair Salons"]
["Cold Stone Creamery"]
["H-E-B","Whataburger","walmart","Stripes Convenience Stores","Valero Energy","Exxon M
["walmart","Sonic","Phillips 66","mcdonalds","Braum's Ice Cream and Dairy Stores","Chi
["Publix Super Markets"]
["Fresenius Medical Care North America"]
["SUBWAY"]
["starbucks","Costco Wholesale Corp.","Chevron","Safeway","Target","Shell Oil","mcdona
["Dunkin' Donuts","ShopRite"]
["starbucks","Safeway","Chevron","Costco Wholesale Corp.","Target","In-N-Out Burger","
["Wawa"]
["Circle K Stores","QuikTrip","walmart","mcdonalds","Fry's Food & Drug Stores","Jack i
["Sheetz"]
["CVS"]
["starbucks","Costco Wholesale Corp.","Chevron","Safeway","Target","Shell Oil","In-N-O
["starbucks","Costco Wholesale Corp.","Target","Chevron","Safeway","mcdonalds","7-Elev
["starbucks","mcdonalds","ARCO","Chevron","In-N-Out Burger","Target","Jack in the Box"
["Casey's General Stores"]
["MasterCuts","Little Caesars"]

["walmart","Exxon Mobil","BI-LO","Chick-fil-A","mcdonalds","Shell Oil","Dollar General
["walmart","QuikTrip","Whataburger","mcdonalds","Jack in the Box","Sonic","Shell Oil"
["walmart","Maverik","Sinclair Oil","mcdonalds","Chevron","Smith's Food & Drug Stores"
["walmart","Sonic","Whataburger","Allsup's","mcdonalds","SUBWAY","Stripes Convenience
["Casey's General Stores","walmart","BP","mcdonalds","SUBWAY","Hy-Vee","Dollar General
["Sonic","walmart","Chevron","Dollar General","Popeyes Louisiana Kitchen","United Stat
["walmart","Dollar General","mcdonalds","Shell Oil","Sonic","SUBWAY","Wendy's","Lowe's
["walmart","mcdonalds","Circle K Stores","Shell Oil","Dollar General","Chick-fil-A","W
["Sheetz","mcdonalds","walmart","SUBWAY","Exxon Mobil","Shell Oil","Chick-fil-A","7-El
["walmart","QuikTrip","Whataburger","mcdonalds","Kroger","Chick-fil-A","Sonic","starbu
["mcdonalds","walmart","Marathon Petroleum","Kroger","SUBWAY","Taco Bell","Shell Oil"
["walmart","mcdonalds","ARCO","starbucks","Jack in the Box","Chevron","Carl's Jr.","Co
["mcdonalds","Walgreens","Speedway","walmart","Family Dollar Stores","Dunkin' Donuts"
["walmart","SUBWAY","Tesoro","Casey's General Stores","Pilot Travel Centers","mcdonald
["Jack in the Box","ARCO","mcdonalds","starbucks","Stater Bros. Markets","Del Taco","T
["Meijer","Speedway","mcdonalds","SUBWAY","BIGGBY COFFEE","Marathon Petroleum","Costco
["mcdonalds","walmart","Dollar General","Speedway","Taco Bell","Kroger","Casey's Gener
["Circle K Stores","walmart","Fry's Food & Drug Stores","QuikTrip","mcdonalds","Shell
["starbucks","Chevron","mcdonalds","Costco Wholesale Corp.","walmart","Target","Safewa
["Whataburger","H-E-B","Stripes Convenience Stores","walmart","Valero Energy","Sunoco"
["walmart","mcdonalds","Dollar General","Shell Oil","Chick-fil-A","SUBWAY","Sonic","Ci
["Dunkin' Donuts","starbucks","SUBWAY","Baskin Robbins","mcdonalds","BP","Rite Aid","T
["Dunkin' Donuts","Hannaford Supermarkets","Irving Oil","Cumberland Farms","Market Bas
```

```
["starbucks","walmart","Sonic","mcdonalds","Phillips 66","Casey's General Stores","Kw
["Dunkin' Donuts","Stop & Shop","CVS","Cumberland Farms","walmart","Mobil","T.J. Maxx
["H-E-B","walmart","Whataburger","Exxon Mobil","Valero Energy","Shell Oil","Sonic","Bo
["starbucks","mcdonalds","SUBWAY","ARCO","Chevron","7-Eleven US","Target","In-N-Out Bu
["mcdonalds","walmart","Pilot Travel Centers","Dollar General","SUBWAY","Casey's Gener
["Kwik Trip","walmart","mcdonalds","Hy-Vee","Casey's General Stores","Applebee's","Ca
["mcdonalds","Meijer","Shell Oil","SUBWAY","walmart","Auto Wares","Speedway","Marathon
Name: related_same_month_brand, Length: 185558, dtype: int64
```

In [12]:
```
data_df['top_brands'].value_counts()
```

Out[12]:
```
[]
["United States Postal Service (USPS)"]
["Aflac (American Family Life Assurance)"]
["Dollar General"]
["National Association for the Education of Young Children (NAEYC)"]
["Cricket Wireless"]
["The American Legion"]
["Shell Oil"]
["Boys & Girls Clubs of America"]
["7-Eleven US"]
["CITGO Petroleum"]
["VFW (Veterans of Foreign Wars)"]
["Family Dollar Stores"]
["Marathon Petroleum"]
["BP"]
["Sunoco"]
["Chevron"]
["United States Postal Service (USPS)","The American Legion"]
["Dollar General","United States Postal Service (USPS)"]
["SUBWAY"]
["Auto Wares"]
["State Farm"]
["Exxon Mobil"]
["Valero Energy"]
["Mobil"]
["Walgreens"]
["CrossFit"]
["CVS"]
["YMCA"]
["Dunkin' Donuts"]

["Chick-fil-A","Walgreens","CVS","Chevrolet","LabCorp","MinuteClinic"]
["Dollar General","Pizza Hut","United States Postal Service (USPS)"]
["American Job Centers","New York State DMV","Kinney Drugs","Sunoco","Do It Best","Ber
["Chick-fil-A","Einstein Brothers","Colton's Steakhouse","SUBWAY","Panera Bread","Au B
["mcdonalds","Chevron","KFC US","starbucks","Candlewood Suites","Jimmy John's","Americ
["starbucks","Society of St. Vincent de Paul","mcdonalds","Country Inn & Suites By Ca
```

```
["AutoZone","BP","Fastenal","Airgas","O'Reilly Auto Parts","Trane","Napa Auto Parts","
["Chevron","Taco Bell","Family Dollar Stores","Walgreens","Americas Best Value Inn","A
["Mobil","Circle K Stores","VFW (Veterans of Foreign Wars)","Enterprise Rent-A-Car"]
["Shell Oil","The Huntington National Bank","Mathnasium"]
["walmart","Hobby Lobby Stores","Hickory Farms","Amigos/Kings Classic","Burger King US
["Pollo Campero","The Counter"]
["Suzuki Cycles","Buick","Chevrolet","GMC (General Motors Company)","Flooring America'
["Ricker's","CBA"]
["Dickey's Barbecue Pit","Discount Drug Mart","Kettering Health Network","Donatos Pizz
["Chick-fil-A","Walgreens","BP","Extra Space Storage","Rent A Wheel","Public Storage"
["Mobil","CITGO Petroleum","United States Postal Service (USPS)","Charter Communicatio
["Shell Oil","VFW (Veterans of Foreign Wars)","Cricket Wireless"]
["walmart","Sam's Club","Lowe's","T.J. Maxx","AMC Entertainment","IMAX","Hobby Lobby S
["Carroll Tire","Cricket Wireless"]
["Shell Oil","Boost Mobile","ConocoPhillips"]
["mcdonalds","Dunkin' Donuts","Dollar Tree","Allstate Insurance","Coldwell Banker","St
["CrossFit","Marathon Petroleum","SUBWAY","Circle K Stores","Dollar General","Century
["Keva Juice","Alphagraphics","Health Mart"]
["CVS","O'Reilly Auto Parts","CHRISTUS Health","DaVita"]
["Rite Aid","Domino's Pizza"]
["Randstad","Ross Stores","QuikTrip","ULTA Beauty","Famous Footwear","Toyota","Z-Spray
["Jack's Family Restaurants","RaceWay","Vermeer","Chevron"]
["Pilot Travel Centers","Taco Bell","SUBWAY"]
["mcdonalds","Walgreens","LabCorp","Marathon Petroleum","United States Postal Service
Name: top_brands, Length: 98086, dtype: int64
```

In [13]:
```
number_data = ['date_range_start','date_range_end','raw_visit_count','raw_visitor_cou
data_df[number_data].describe()
```

Out[13]:

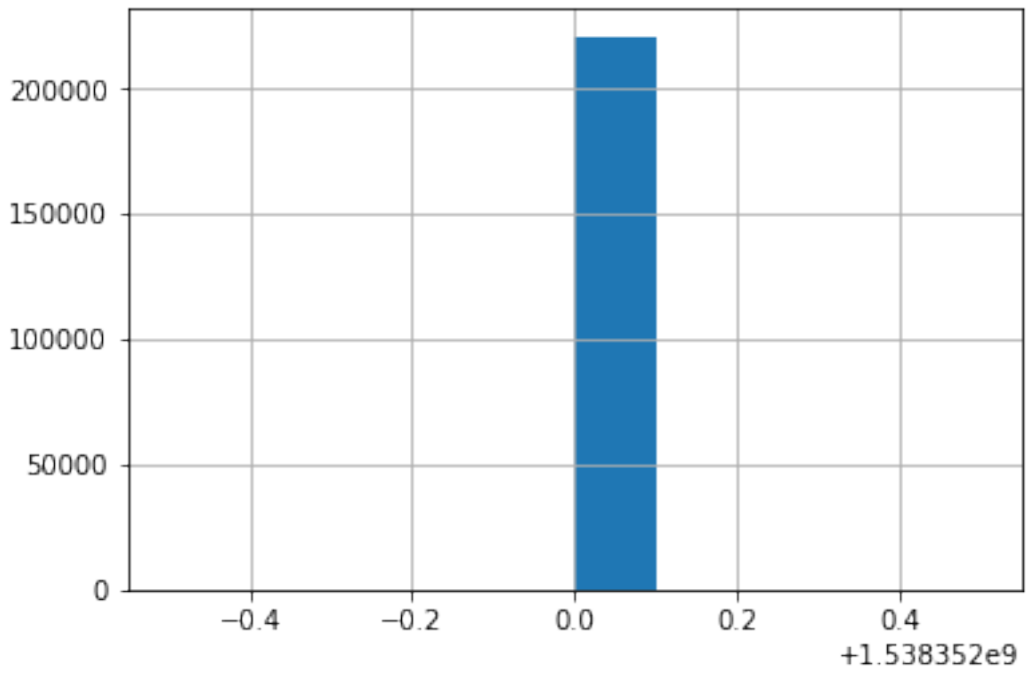
	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	\
count	2.207350e+05	2.207350e+05	2.206290e+05	2.206290e+05	
mean	1.538352e+09	1.541030e+09	4.793066e+04	1.182032e+04	
std	0.000000e+00	0.000000e+00	6.252655e+04	3.045832e+04	
min	1.538352e+09	1.541030e+09	6.000000e+01	5.000000e+01	
25%	1.538352e+09	1.541030e+09	1.704200e+04	3.430000e+03	
50%	1.538352e+09	1.541030e+09	3.064000e+04	6.541000e+03	
75%	1.538352e+09	1.541030e+09	5.667800e+04	1.309900e+04	
max	1.538352e+09	1.541030e+09	7.179900e+06	6.113949e+06	

	distance_from_home
count	2.205180e+05
mean	3.511280e+04
std	9.973193e+04
min	7.060000e+02
25%	8.584000e+03
50%	1.461400e+04
75%	3.139775e+04
max	6.297845e+06

0.2

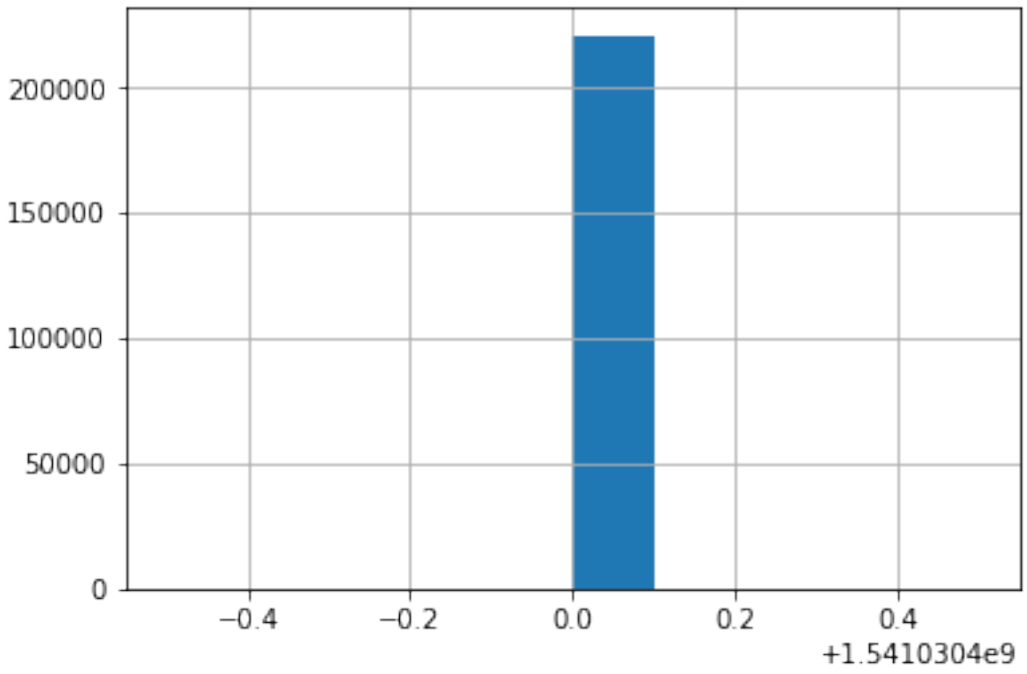
```
In [19]: data_df['date_range_start'].hist()
```

```
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x1c30b462860>
```



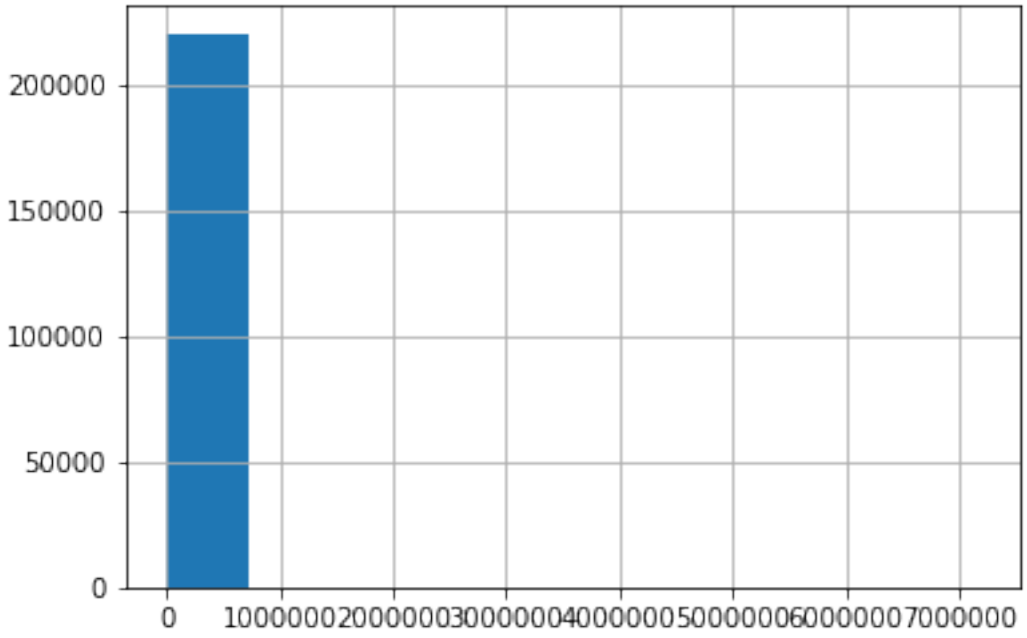

```
In [15]: data_df['date_range_end'].hist()
```

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1c30b2d8d68>
```



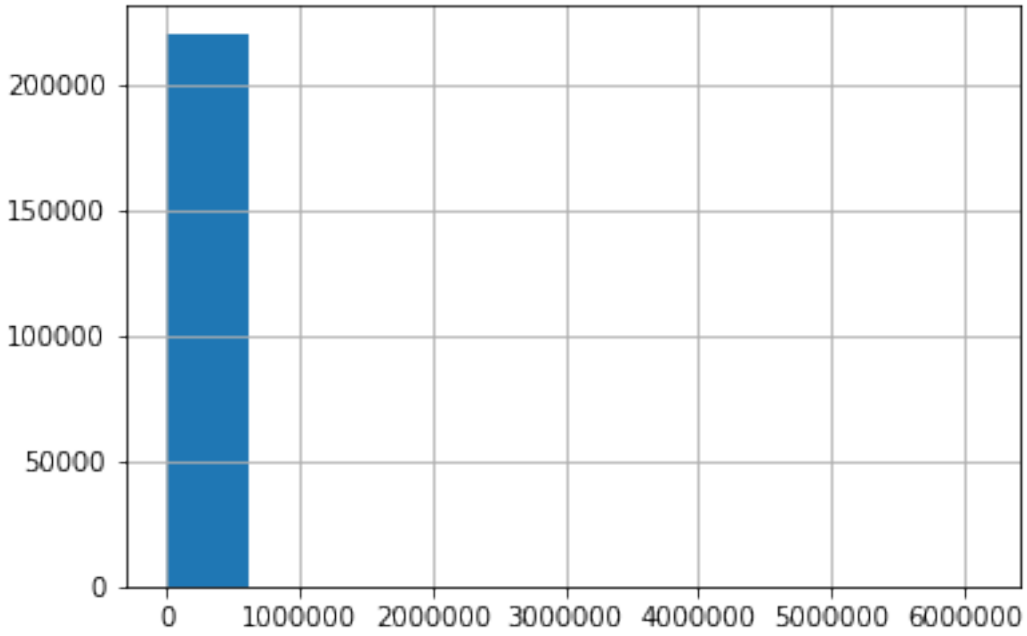

```
In [16]: data_df['raw_visit_count'].hist()
```

```
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x1c30b293dd8>
```



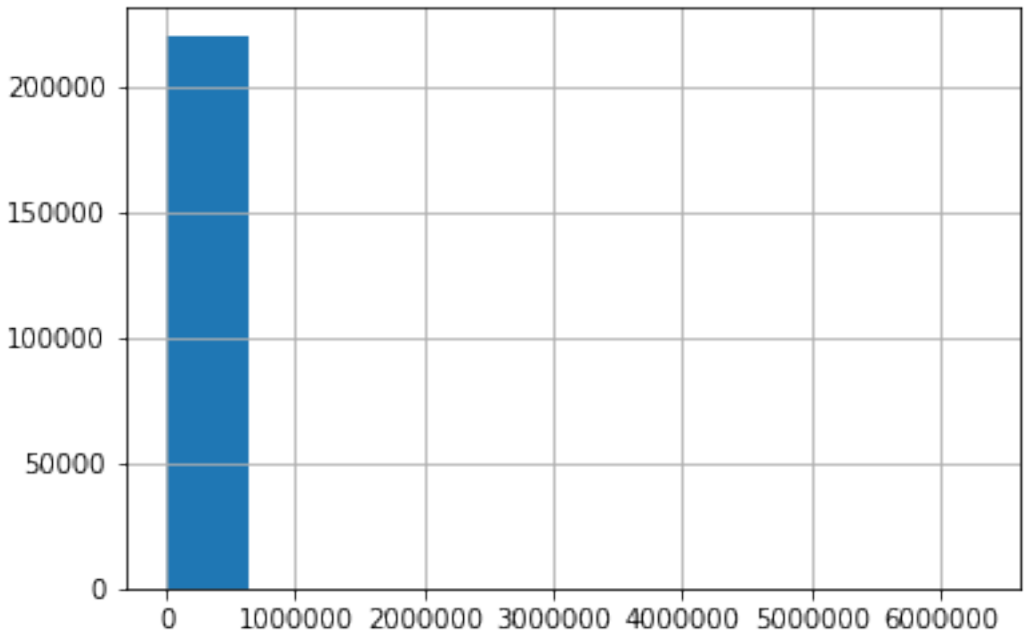
```
In [17]: data_df['raw_visitor_count'].hist()
```

```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x1c30b01f9b0>
```




```
In [18]: data_df['distance_from_home'].hist()
```

```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x1c30b2bd278>
```

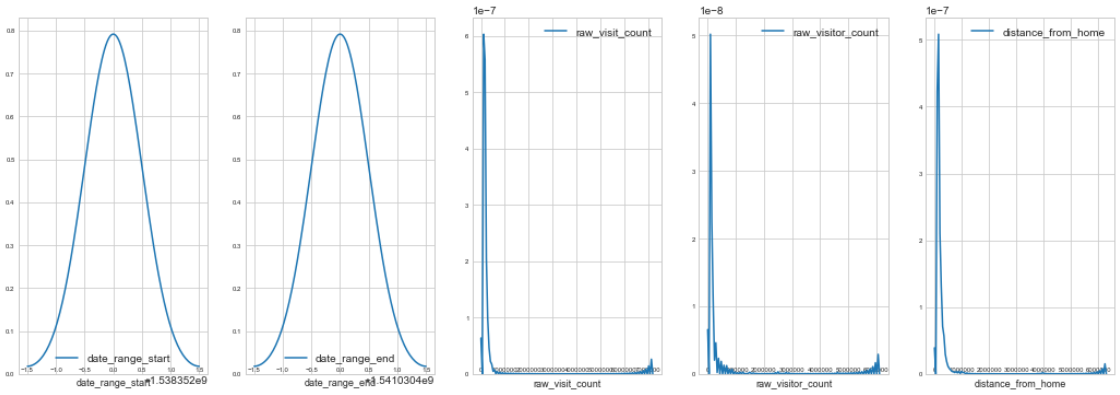


In [22]:
```
import matplotlib.pyplot as plt
import seaborn as sns
def plot_feature_distribution(df, features):
    i = 0
    sns.set_style('whitegrid')
    plt.figure()
    fig, ax = plt.subplots(1,5,figsize=(18,6))

    for feature in features:
        i += 1
        plt.subplot(1,5,i)
        sns.kdeplot(df[feature], bw=0.5,label=feature)
        plt.xlabel(feature, fontsize=9)
        locs, labels = plt.xticks()
        plt.tick_params(axis='x', which='major', labelsize=6, pad=-6)
        plt.tick_params(axis='y', which='major', labelsize=6)
    plt.show();
```

In [23]:
```
features = ['date_range_start','date_range_end','raw_visit_count','raw_visitor_count'
plot_feature_distribution(data_df, features)
```

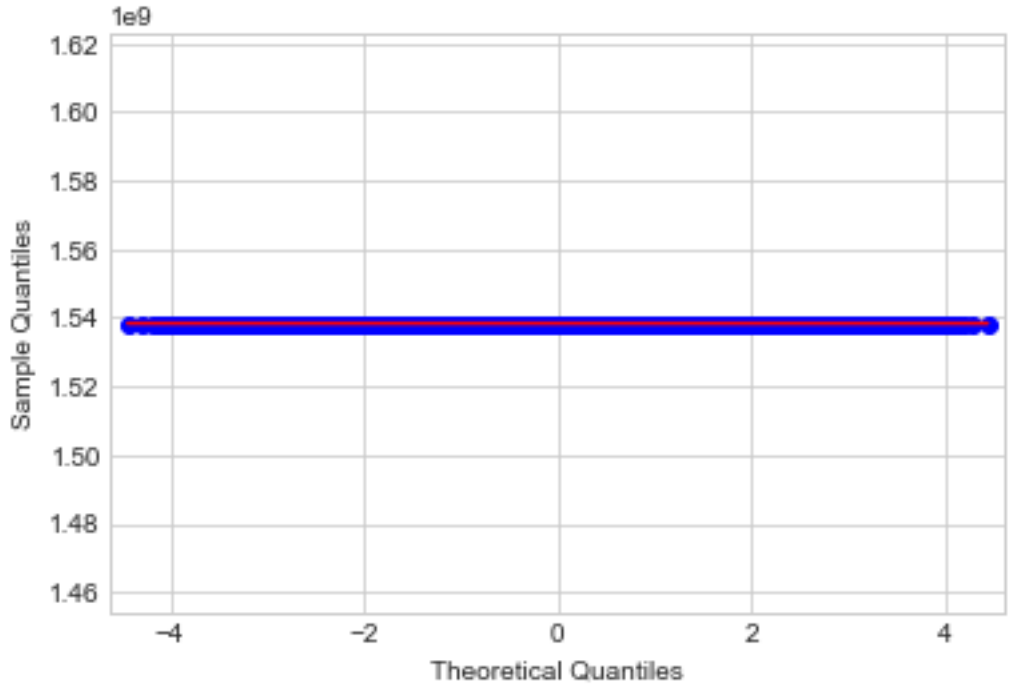
```
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\nonparametric\kde.py:448: RuntimeWarning
  X = X[np.logical_and(X > clip[0], X < clip[1])] # won't work for two columns.
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\nonparametric\kde.py:448: RuntimeWarning
  X = X[np.logical_and(X > clip[0], X < clip[1])] # won't work for two columns.
```

```
<Figure size 432x288 with 0 Axes>
```

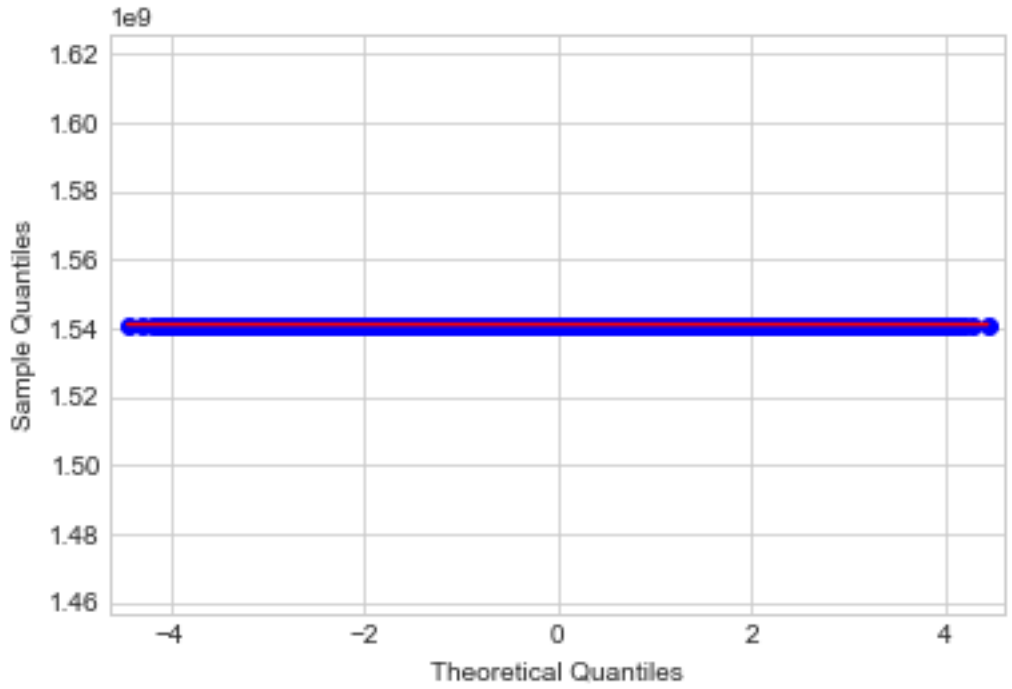


In [24]:
```
import statsmodels.api as sm
import pylab
```

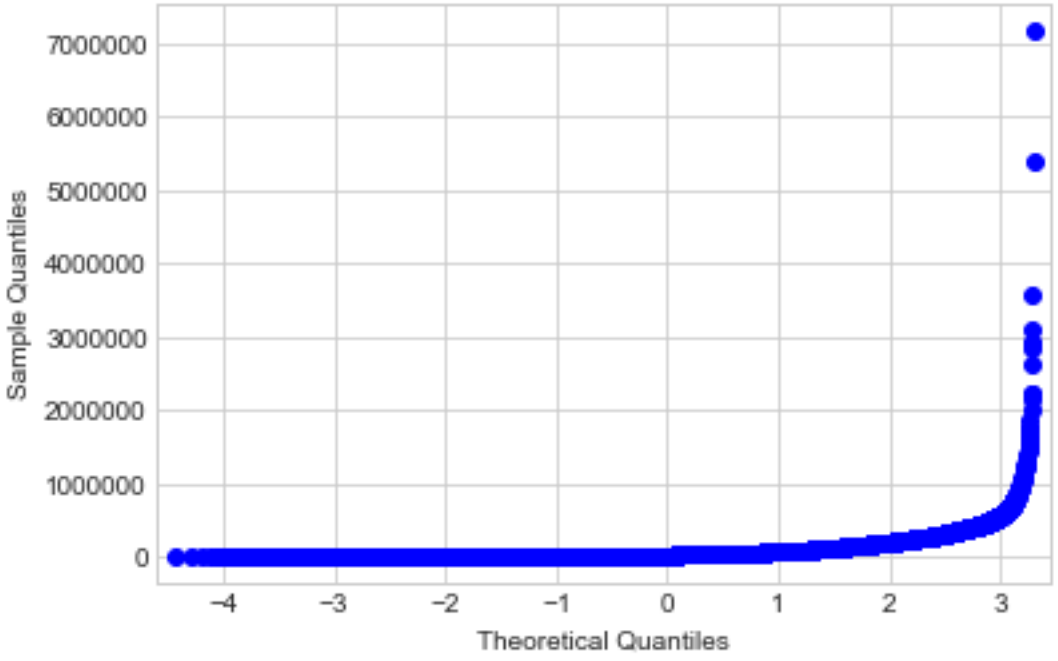
```
sm.qqplot(data_df['date_range_start'], line='s')
pylab.show()
```



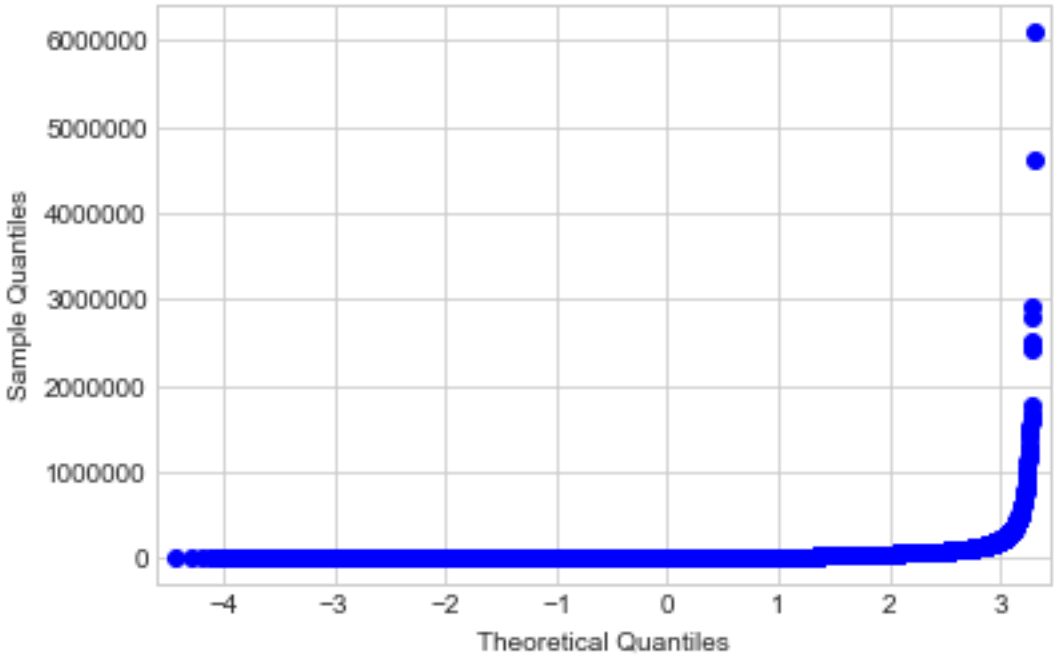
In [25]:
```
sm.qqplot(data_df['date_range_end'], line='s')
pylab.show()
```



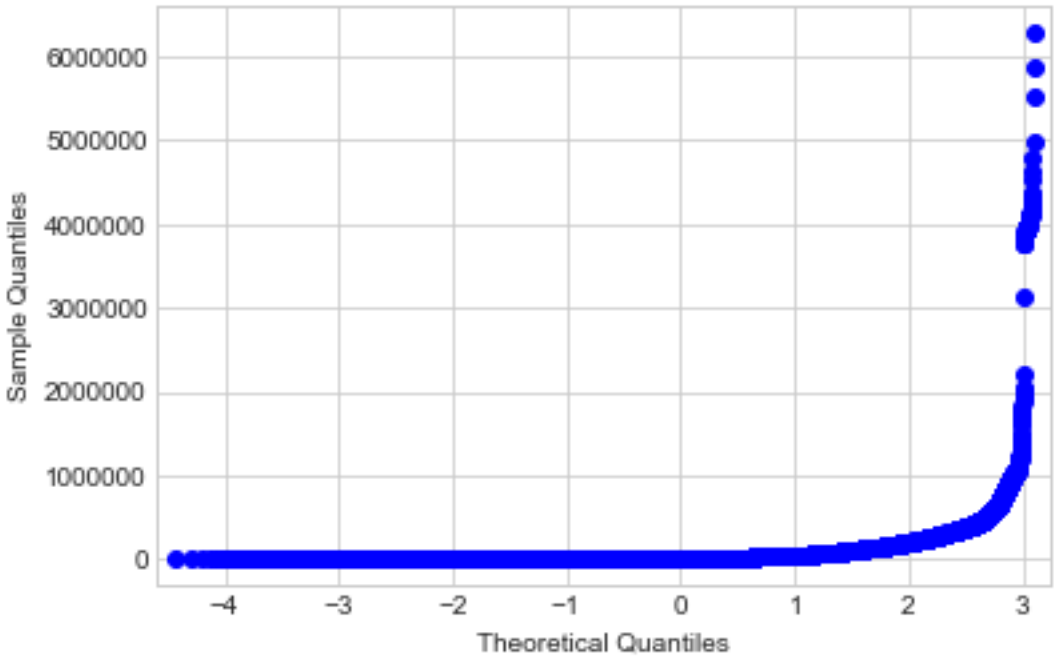
In [26]:
```
sm.qqplot(data_df['raw_visit_count'], line='s')
pylab.show()
```



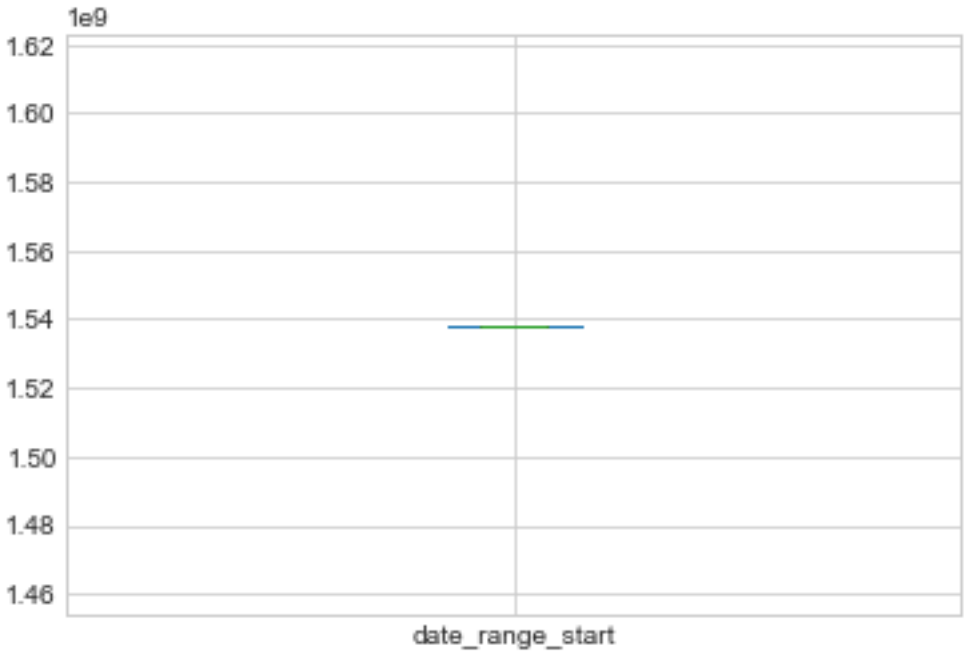
In [27]:
```
sm.qqplot(data_df['raw_visitor_count'], line='s')
pylab.show()
```



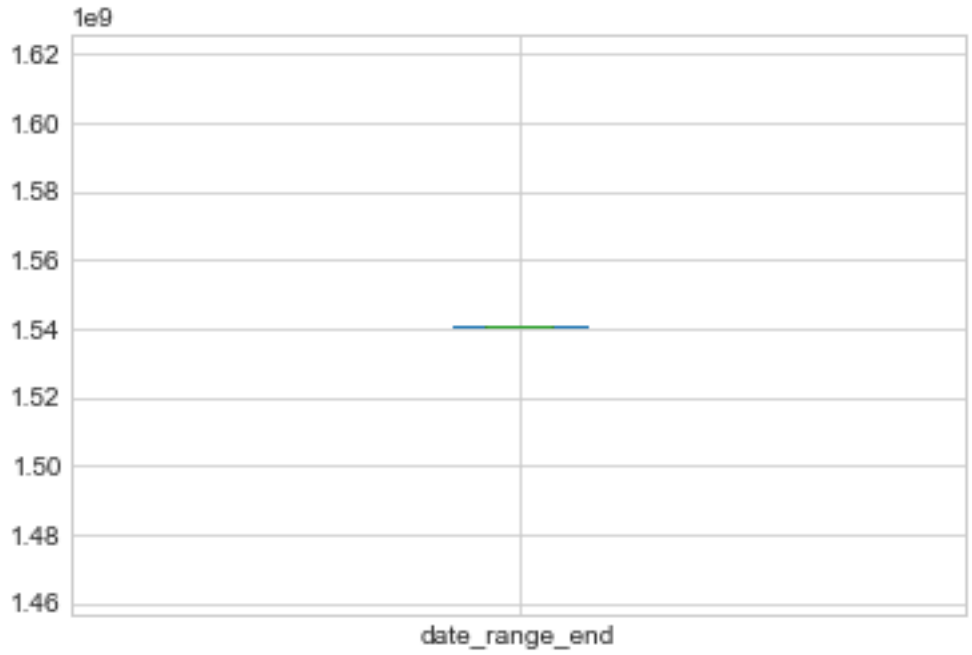
In [28]:
```
sm.qqplot(data_df['distance_from_home'], line='s')
pylab.show()
```



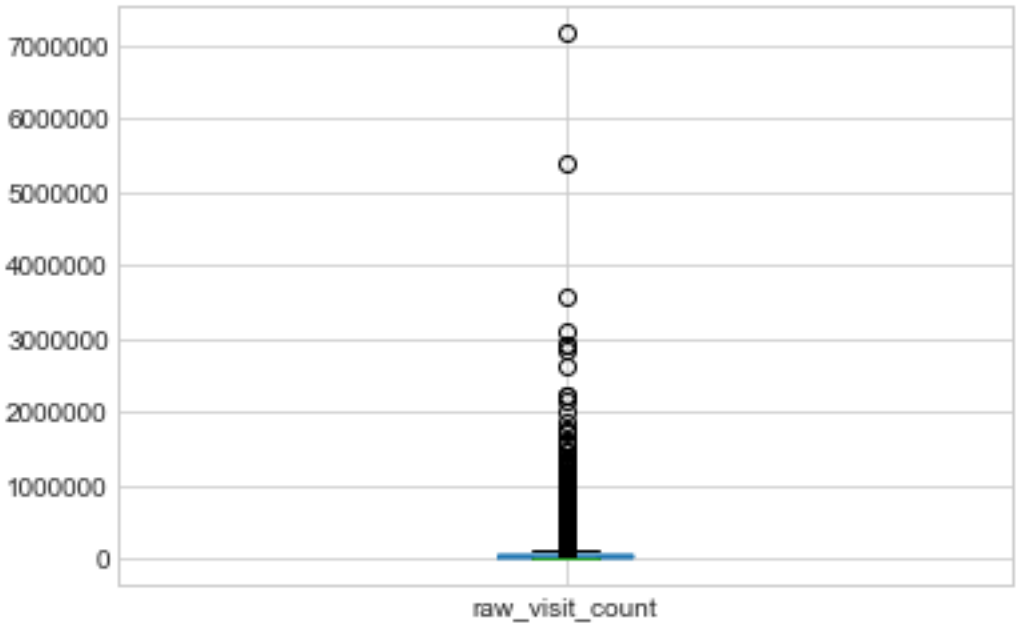
In [29]:
```
import matplotlib.pyplot as plt

data_df['date_range_start'].plot(kind='box', notch=True, grid=True)
plt.show()
```



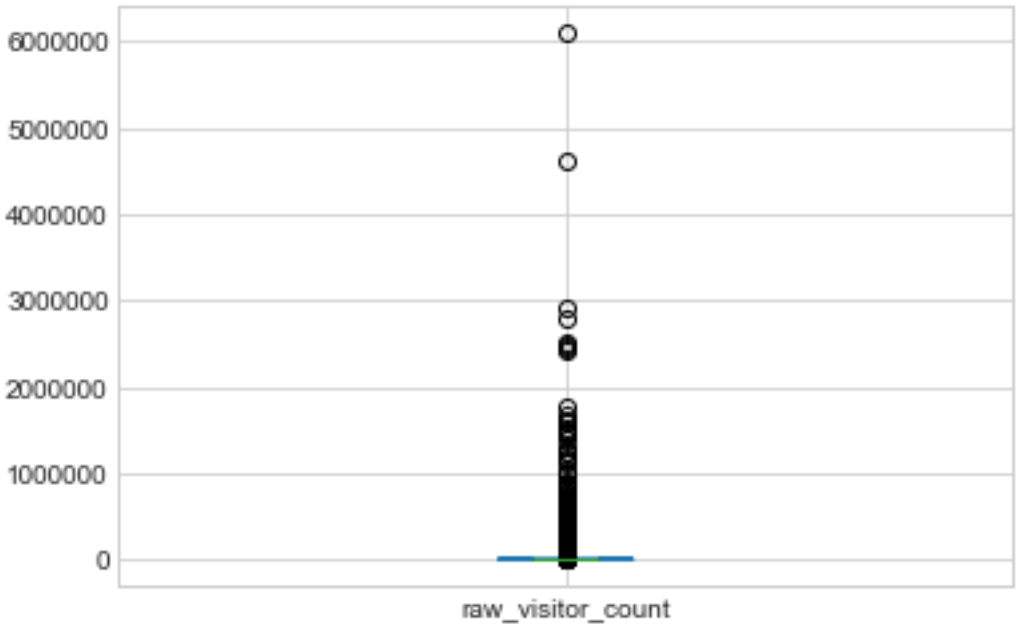
In [30]:
```
data_df['date_range_end'].plot(kind='box', notch=True, grid=True)
plt.show()
```



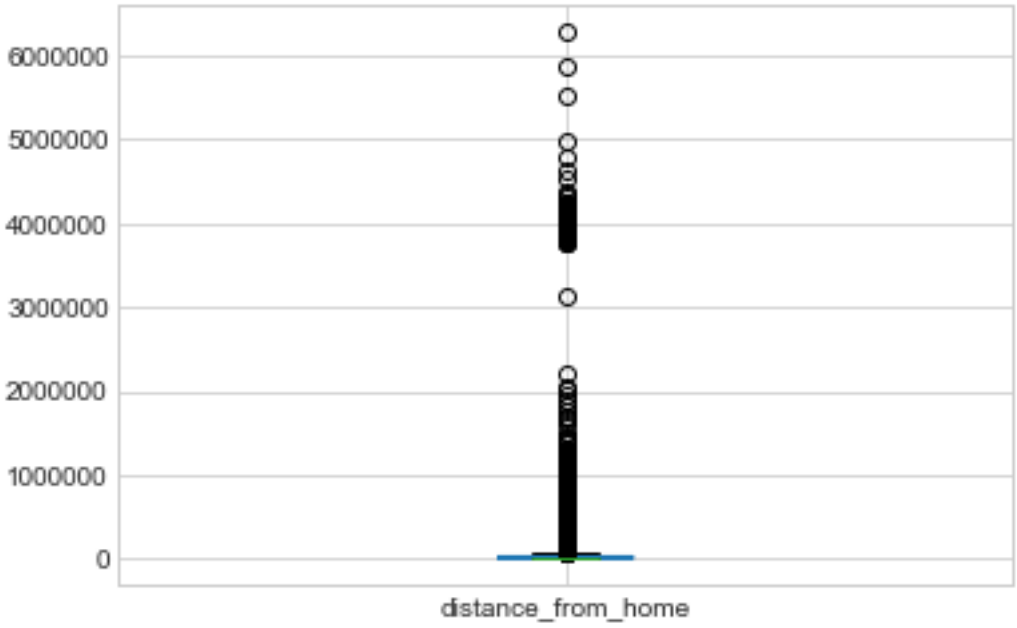
In [31]:
```
data_df['raw_visit_count'].plot(kind='box', notch=True, grid=True)
plt.show()
```



In [32]:
```
data_df['raw_visitor_count'].plot(kind='box', notch=True, grid=True)
plt.show()
```



In [33]:
```
data_df['distance_from_home'].plot(kind='box', notch=True, grid=True)
plt.show()
```



0.3

```
In [34]: def missing_data(data):
             total = data.isnull().sum()
             percent = (data.isnull().sum()/data.isnull().count()*100)
             tt = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
             types = []
             for col in data.columns:
                 dtype = str(data[col].dtype)
                 types.append(dtype)
             tt['Types'] = types
             return(np.transpose(tt))

In [35]: %%time
         missing_data(data_df)

Wall time: 634 ms
```

Out[35]:

	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	\
Total	0	0	106	106	
Percent	0	0	0.0480214	0.0480214	
Types	int64	int64	float64	float64	

	visitor_home_cbgs	visitor_work_cbgs	distance_from_home	\
Total	0	0	217	
Percent	0	0	0.0983079	
Types	object	object	float64	

	related_same_day_brand	related_same_month_brand	top_brands	\
Total	0	0	0	
Percent	0	0	0	
Types	object	object	object	

	popularity_by_hour	popularity_by_day
Total	0	0
Percent	0	0
Types	object	object

0.3.1

```
In [36]: del_df = data_df.dropna()

In [37]: del_df.shape

Out[37]: (220518, 12)

In [38]: del_df.head()
```

Out[38]:

census_block_group	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	visitor_home_cbgs	visitor_work_cbgs	distance_from_home	related_same_day_brand
1.005951e+10	1538352000	1541030400	75122.0	18314.0	{"010059501003":127,"010059509001":111,"010059...	{"010059501003":109,"010810407002":62,"0108104...	194724.0	["Chick-fil-A","mcdonalds","Marathon Petroleum...
1.009051e+10	1538352000	1541030400	95649.0	38942.0	{"010730113021":210,"010090506022":205,"010090...	{"010890111001":271,"010730045001":269,"010439...	120587.0	["Shell Oil","mcdonalds","Chick-fil-A","Chevron"]
1.047957e+10	1538352000	1541030400	14009.0	3039.0	{"010479567011":67,"010479567021":60}	{"010479567021":52}	67774.0	["Dollar General"]
1.069040e+10	1538352000	1541030400	128169.0	25418.0	{"010690402013":370,"010690402011":322,"010690...	{"010690402024":313,"010690415004":203,"010450...	42684.0	["Chick-fil-A","Sam's Club","Dollar General","...
1.073011e+10	1538352000	1541030400	51453.0	9499.0	{"010090507001":183,"010730113021":167,"010730...	{"010730045001":140,"010730027001":123,"010730...	18878.0	["Chevron","Daylight Donuts","walmart"]

```
                                                    related_same_month_brand  \
census_block_group
1.005951e+10        ["walmart","mcdonalds","Dollar General","Chick...
1.009051e+10        ["walmart","mcdonalds","Shell Oil","Chick-fil-...
1.047957e+10        ["walmart","Dollar General","mcdonalds","Chevr...
1.069040e+10        ["walmart","Dollar General","mcdonalds","Marat...
1.073011e+10        ["walmart","Chevron","Dollar General","Shell O...

                                                          top_brands  \
census_block_group
1.005951e+10             ["CrossFit","Health Mart","Coldwell Banker"]
1.009051e+10                                                       []
1.047957e+10                                          ["Dollar General"]
1.069040e+10        ["Chick-fil-A","Sam's Club","Olive Garden","mc...
1.073011e+10                                     ["Chevron","CrossFit"]

                                                     popularity_by_hour  \
census_block_group
1.005951e+10        [2617,2457,2403,2519,2646,3007,3886,7566,5508,...
1.009051e+10        [6556,6325,6222,6355,6586,7350,8568,8099,7378,...
1.047957e+10        [807,790,796,786,851,951,1134,1797,1355,1241,1...
1.069040e+10        [2121,1828,1784,1704,1861,2373,3730,7497,7093,...
1.073011e+10        [3804,3716,3686,3672,3735,4115,4855,5946,4526,...

                                                      popularity_by_day
census_block_group
1.005951e+10        {"Monday":12000,"Tuesday":12224,"Wednesday":10...
1.009051e+10        {"Monday":12125,"Tuesday":12984,"Wednesday":12...
1.047957e+10        {"Monday":2314,"Tuesday":2340,"Wednesday":2195...
1.069040e+10        {"Monday":21141,"Tuesday":21143,"Wednesday":17...
1.073011e+10        {"Monday":8402,"Tuesday":8414,"Wednesday":8550...
```

0.3.2

In [40]:
```
from collections import Counter
from math import isnan

miss_features = ['raw_visit_count','raw_visitor_count','distance_from_home']
fill_df = data_df

for col in miss_features:
    word_counts = Counter(fill_df[col])
    top = word_counts.most_common(1)[0][0]
    if type(top) != str:
        if isnan(top):
            top = word_counts.most_common(2)[1][0]
    print(top, type(top))
    temp = fill_df[col].fillna(top)
```

```
            fill_df[col] = temp
        fill_df.head()
```

```
24531.0 <class 'float'>
2292.0 <class 'float'>
8345.0 <class 'float'>
```

Out[40]:

census_block_group	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	visitor_home_cbgs	visitor_work_cbgs	distance_from_home
1.005951e+10	1538352000	1541030400	75122.0	18314.0	{"010059501003":127,"010059509001":111,"010059...	{"010059501003":109,"010810407002":62,"0108104...	194724.0
1.009051e+10	1538352000	1541030400	95649.0	38942.0	{"010730113021":210,"010090506022":205,"010090...	{"010890111001":271,"010730045001":269,"010439...	120587.0
1.047957e+10	1538352000	1541030400	14009.0	3039.0	{"010479567011":67,"010479567021":60}	{"010479567021":52}	67774.0
1.069040e+10	1538352000	1541030400	128169.0	25418.0	{"010690402013":370,"010690402011":322,"010690...	{"010690402024":313,"010690415004":203,"010450...	42684.0
1.073011e+10	1538352000	1541030400	51453.0	9499.0	{"010090507001":183,"010730113021":167,"010730...	{"010730045001":140,"010730027001":123,"010730...	18878.0

```
                                                    related_same_day_brand  \
census_block_group
1.005951e+10        ["Chick-fil-A","mcdonalds","Marathon Petroleum...
1.009051e+10        ["Shell Oil","mcdonalds","Chick-fil-A","Chevron"]
1.047957e+10                                        ["Dollar General"]
1.069040e+10        ["Chick-fil-A","Sam's Club","Dollar General","...
1.073011e+10                  ["Chevron","Daylight Donuts","walmart"]

                                                   related_same_month_brand  \
census_block_group
1.005951e+10        ["walmart","mcdonalds","Dollar General","Chick...
1.009051e+10        ["walmart","mcdonalds","Shell Oil","Chick-fil-...
1.047957e+10        ["walmart","Dollar General","mcdonalds","Chevr...
1.069040e+10        ["walmart","Dollar General","mcdonalds","Marat...
1.073011e+10        ["walmart","Chevron","Dollar General","Shell O...

                                                             top_brands  \
census_block_group
1.005951e+10             ["CrossFit","Health Mart","Coldwell Banker"]
1.009051e+10                                                        []
1.047957e+10                                        ["Dollar General"]
1.069040e+10        ["Chick-fil-A","Sam's Club","Olive Garden","mc...
1.073011e+10                                   ["Chevron","CrossFit"]

                                                    popularity_by_hour  \
census_block_group
1.005951e+10        [2617,2457,2403,2519,2646,3007,3886,7566,5508,...
1.009051e+10        [6556,6325,6222,6355,6586,7350,8568,8099,7378,...
1.047957e+10        [807,790,796,786,851,951,1134,1797,1355,1241,1...
1.069040e+10        [2121,1828,1784,1704,1861,2373,3730,7497,7093,...
1.073011e+10        [3804,3716,3686,3672,3735,4115,4855,5946,4526,...

                                                     popularity_by_day
census_block_group
1.005951e+10        {"Monday":12000,"Tuesday":12224,"Wednesday":10...
1.009051e+10        {"Monday":12125,"Tuesday":12984,"Wednesday":12...
1.047957e+10        {"Monday":2314,"Tuesday":2340,"Wednesday":2195...
1.069040e+10        {"Monday":21141,"Tuesday":21143,"Wednesday":17...
1.073011e+10        {"Monday":8402,"Tuesday":8414,"Wednesday":8550...
```

In [41]:
```
missing_data(fill_df)
```

Out[41]:
```
         date_range_start date_range_end raw_visit_count raw_visitor_count  \
Total                   0              0               0                  0
Percent                 0              0               0                  0
Types               int64          int64         float64            float64

         visitor_home_cbgs visitor_work_cbgs distance_from_home  \
```

```
Total                   0                 0                  0
Percent                 0                 0                  0
Types              object            object            float64

        related_same_day_brand related_same_month_brand top_brands  \
Total                        0                       0          0
Percent                      0                       0          0
Types                   object                  object     object

        popularity_by_hour popularity_by_day
Total                    0                 0
Percent                  0                 0
Types               object            object
```

0.3.3

In [43]:
```
data_df.corr()
```

Out[43]:

	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	distance_from_home
date_range_start	NaN	NaN	NaN	NaN	NaN
date_range_end	NaN	NaN	NaN	NaN	NaN
raw_visit_count	NaN	NaN	1.000000	0.815860	0.021867
raw_visitor_count	NaN	NaN	0.815860	1.000000	0.032968
distance_from_home	NaN	NaN	0.021867	0.032968	1.000000

In [44]:
```
data_df.corr('spearman')
```

Out[44]:

	date_range_start	date_range_end	raw_visit_count	raw_visitor_count	distance_from_home
date_range_start	NaN	NaN	NaN	NaN	NaN
date_range_end	NaN	NaN	NaN	NaN	NaN
raw_visit_count	NaN	NaN	1.000000	0.937175	0.182710
raw_visitor_count	NaN	NaN	0.937175	1.000000	0.199526
distance_from_home	NaN	NaN	0.182710	0.199526	1.000000

0.3.4

```
In [ ]: from fancyimpute import KNN

        features = ['date_range_start','date_range_end','raw_visit_count','raw_visitor_count',
        train_data_x = data_df[features]
        train_data_x = pd.DataFrame(KNN(k=6).fit_transform(train_data_x), columns=features)
```