Exploratory Data Analysis

In [94]:
```
#Importing Libraries
import numpy as np
import pandas as pd
import seaborn as sns
sns.set_style('whitegrid')
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
df = pd.DataFrame()
```

In [10]:
```
#Reading the clean data
df = pd.read_excel('Stack_Overflow_Questions_Clean_Data.xlsx')
```

Let's explore data through pandas

In [12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20000 entries, 0 to 19999
Data columns (total 12 columns):
Question Id           20000 non-null int64
Votes                 20000 non-null int64
Answer Count          5710 non-null float64
Views                 20000 non-null int64
Question              20000 non-null object
QDescription          20000 non-null object
User                  19953 non-null object
Reputation Score      19953 non-null float64
Gold Badge Count      2296 non-null float64
Silver Badge Count    5400 non-null float64
Bronze Badge Count    18388 non-null float64
Tags                  20000 non-null object
dtypes: float64(5), int64(3), object(4)
memory usage: 2.0+ MB
```

In [64]:
```
df.columns
```

Out[64]:
```
Index(['Question Id', 'Votes', 'Answer Count', 'Views', 'Question',
       'QDescription', 'User', 'Reputation Score', 'Gold Badge Count',
       'Silver Badge Count', 'Bronze Badge Count', 'Tags'],
      dtype='object')
```

In [14]:
```
df.shape
```

Out[14]:
```
(20000, 12)
```

In [17]:
```
df.head()
```

Out[17]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count
0	49369882	0	NaN	3	Tag items with color	I have two squares in my PyOpenGL window insid...	BjkOcean	29.0	NaN
1	49369867	0	NaN	5	Firebase dump json data	I'm no back-end developer. So perspective is a...	user9132502	52.0	NaN
2	49369855	-5	NaN	14	Specific type of webscraping [on hold]	How can I get python to check a specific line ...	Ronprogramming	4.0	NaN
3	49369846	0	NaN	3	Notification on android via Python using Pushs...	I am using "Pushsafer" to notify on my smartph...	Sanket	1.0	NaN
4	49369842	0	NaN	14	How to get the exact count of people by face d...	I am working on getting the total count of peo...	Mueez Siraj	101.0	NaN

In [18]:
```
df.tail()
```

Out[18]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputati Sc
19995	48888000	1	NaN	25	python mySQLdb module not found issue in windows	I want to retrieve data from mysql using pytho...	Codemaker	3
19996	48887991	2	NaN	36	Rolling data based on custom function - Pandas	I am trying to create a DataFrame that contain...	robgreen48	2
19997	48887986	-1	1.0	32	The usage of np.random.seed when we change its...	According to What does numpy.random.seed(0) do...	David	6
19998	48887954	0	NaN	26	Reading mouse movements with Python	I need to move the mouse to x and y coordinate...	hurkaperpa	6
19999	48887912	5	NaN	98	Find minimum distance between points of two li...	I have two lists of coordinates:\n\ns1 = [(0,0...	orak	109

In [19]:
```
df.describe()
```

Out[19]:

	Question Id	Votes	Answer Count	Views	Reputation Score	Gold Badge Count	E
count	2.000000e+04	20000.000000	5710.000000	20000.000000	19953.000000	2296.000000	5400.0
mean	4.913211e+07	-0.002800	1.424168	35.452300	569.938505	5.264373	4.0
std	1.387679e+05	1.689981	0.794387	78.155024	5643.715947	9.423064	6.6
min	4.888791e+07	-15.000000	1.000000	2.000000	1.000000	1.000000	1.0
25%	4.901232e+07	0.000000	1.000000	22.000000	4.000000	1.000000	1.0
50%	4.913367e+07	0.000000	1.000000	30.000000	20.000000	2.000000	2.0
75%	4.924947e+07	1.000000	2.000000	41.000000	119.000000	5.000000	4.0
max	4.936988e+07	64.000000	8.000000	6000.000000	486000.000000	96.000000	96.0

Exploring Missing Values

In [20]:
```
# Exercise 3.2. Find missing values

# determine which value is null
df.isnull()
```

Out[20]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count	Silver Badge Count
0	False	False	True	False	False	False	False	False	True	True
1	False	False	True	False	False	False	False	False	True	True
2	False	False	True	False	False	False	False	False	True	True
3	False	False	True	False	False	False	False	False	True	True
4	False	False	True	False	False	False	False	False	True	True
5	False	False	True	False	False	False	False	False	True	True
6	False	False	True	False	False	False	False	False	True	True
7	False	False	True	False	False	False	False	False	True	False
8	False	False	True	False	False	False	False	False	True	True
9	False	False	True	False	False	False	False	False	True	False
10	False	False	True	False	False	False	False	False	True	True
11	False	False	True	False	False	False	False	False	True	True
12	False	False	True	False	False	False	False	False	True	False
13	False	False	False	False	False	False	False	False	True	True
14	False	False	True	False	False	False	False	False	True	True
15	False	False	True	False	False	False	False	False	True	False
16	False	False	False	False	False	False	False	False	False	False
17	False	False	True	False	False	False	False	False	True	True
18	False	False	True	False	False	False	False	False	True	True
19	False	False	True	False	False	False	False	False	True	True
20	False	False	True	False	False	False	False	False	True	True
21	False	False	True	False	False	False	False	False	True	True
22	False	False	True	False	False	False	False	False	True	True
23	False	False	False	False	False	False	False	False	False	False
24	False	False	True	False	False	False	False	False	True	True
25	False	False	False	False	False	False	False	False	True	True
26	False	False	True	False	False	False	False	False	True	True
27	False	False	False	False	False	False	False	False	True	False
28	False	False	True	False	False	False	False	False	False	False
29	False	False	True	False	False	False	False	False	True	True

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count	Silver Badge Count
...	...	...	...	...	...	...	...	...	...	...
19970	False	False	True	False	False	False	False	False	True	True
19971	False	False	False	False	False	False	False	False	True	True
19972	False	False	True	False	False	False	False	False	True	True
19973	False	False	True	False	False	False	False	False	True	True
19974	False	False	True	False	False	False	False	False	True	True
19975	False	False	True	False	False	False	False	False	True	True
19976	False	False	True	False	False	False	False	False	True	True
19977	False	False	True	False	False	False	False	False	True	True
19978	False	False	True	False	False	False	False	False	True	True
19979	False	False	True	False	False	False	False	False	True	True
19980	False	False	True	False	False	False	False	False	True	True
19981	False	False	False	False	False	False	False	False	True	True
19982	False	False	True	False	False	False	False	False	True	True
19983	False	False	True	False	False	False	False	False	True	True
19984	False	False	True	False	False	False	False	False	True	True
19985	False	False	True	False	False	False	False	False	True	True
19986	False	False	True	False	False	False	False	False	True	False
19987	False	False	False	False	False	False	False	False	True	False
19988	False	False	True	False	False	False	False	False	True	True
19989	False	False	True	False	False	False	False	False	True	True
19990	False	False	True	False	False	False	False	False	True	True
19991	False	False	True	False	False	False	False	False	True	True
19992	False	False	True	False	False	False	False	False	True	True
19993	False	False	False	False	False	False	False	False	True	False
19994	False	False	True	False	False	False	False	False	True	True
19995	False	False	True	False	False	False	False	False	True	True
19996	False	False	True	False	False	False	False	False	True	True
19997	False	False	False	False	False	False	False	False	True	True
19998	False	False	True	False	False	False	False	False	True	True
19999	False	False	True	False	False	False	False	False	False	False

20000 rows × 12 columns

In [21]:
```
df[df.isnull().any(axis=1)]
```

Out[21]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	
0	49369882	0	NaN	3	Tag items with color	I have two squares in my PyOpenGL window insid...	
1	49369867	0	NaN	5	Firebase dump json data	I'm no back-end developer. So perspective is a...	use
2	49369855	-5	NaN	14	Specific type of webscraping [on hold]	How can I get python to check a specific line ...	Ronpro
3	49369846	0	NaN	3	Notification on android via Python using Pushs...	I am using "Pushsafer" to notify on my smartph...	
					How to get the	I am working on	

In [23]:
```
# get number of null values in each column
df.isnull().sum(axis=0)
```

Out[23]:
```
Question Id              0
Votes                    0
Answer Count         14290
Views                    0
Question                 0
QDescription             0
User                    47
Reputation Score        47
Gold Badge Count     17704
Silver Badge Count   14600
Bronze Badge Count    1612
Tags                     0
dtype: int64
```

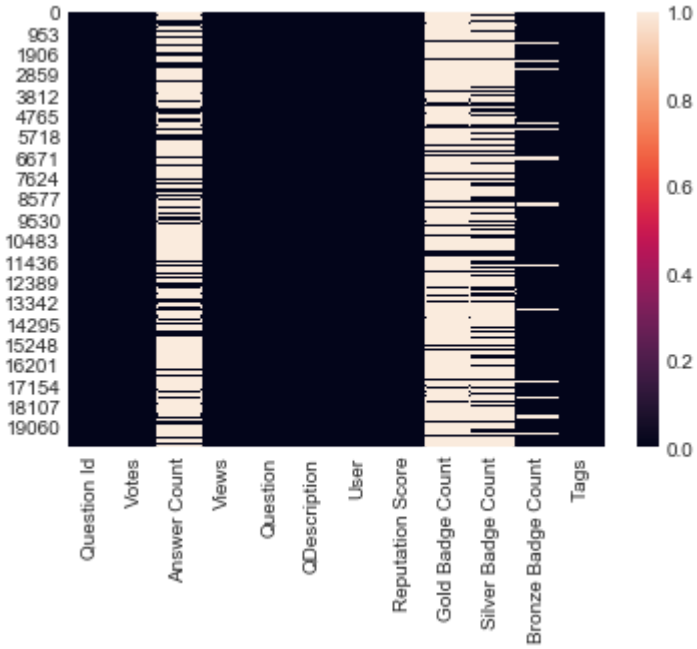
Data Analysis

The data looks good for certain columns but not for all columns. We have 5710 answers for 20000 questions. For user and reputation score, there are roughly around 50 missing values. Which we can impute use scikit learn by imputing median values. And for badges there are lots of missing values.

Let's carry out some Visualization to analyze missing data

In [25]:
```
#Creating heatmap to find out missing columns and scale of missing data
sns.heatmap(df.isna())
```

Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x203b4baf898>



Cleaning Data

Let's replace NaN answercount values with 0

In [27]:
```
df['Answer Count'].fillna(0, inplace=True)
```

In [28]:
```
df.isnull().sum(axis=0)
```

Out[28]:
```
Question Id              0
Votes                    0
Answer Count             0
Views                    0
Question                 0
QDescription             0
User                    47
Reputation Score        47
Gold Badge Count     17704
Silver Badge Count   14600
Bronze Badge Count    1612
Tags                     0
dtype: int64
```

Let's replace NaN Badge Count Values for the three badges Gold, Silver, Bronze values with 0

In [29]:
```
df['Gold Badge Count'].fillna(0, inplace=True)
```

In [30]:
```
df['Silver Badge Count'].fillna(0, inplace=True)
```

In [32]:
```
df['Bronze Badge Count'].fillna(0, inplace=True)
```

Checking the Null Values Count to Verify

In [29]:
```
df.isnull().sum(axis=0)
```

Out[29]:
```
Question Id            0
Votes                  0
Answer Count           0
Views                  0
Question               0
QDescription           0
User                  47
Reputation Score      47
Gold Badge Count       0
Silver Badge Count     0
Bronze Badge Count     0
Tags                   0
dtype: int64
```

Lets interpolate Reputation Score Values

In [34]:
```
df['Reputation Score'].describe()
```

Out[34]:
```
count     19953.000000
mean        569.938505
std        5643.715947
min           1.000000
25%           4.000000
50%          20.000000
75%         119.000000
max      486000.000000
Name: Reputation Score, dtype: float64
```

In [35]:
```
type(df['Reputation Score'])
```

Out[35]:
```
pandas.core.series.Series
```

In [36]:
```
#Let's impute the Missing Values for Reputation Score
```

In [37]:
```
from sklearn.preprocessing import Imputer
```

In [38]:
```
imp = Imputer(missing_values='NaN',strategy= 'median',axis=0)
```

In [39]:
```
imp=Imputer(missing_values="NaN", strategy="median",axis=0)
imp.fit(df[["Reputation Score"]])
df["Reputation Score"]=imp.transform(df[["Reputation Score"]]).ravel()
```

In [40]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20000 entries, 0 to 19999
Data columns (total 12 columns):
Question Id           20000 non-null int64
Votes                 20000 non-null int64
Answer Count          20000 non-null float64
Views                 20000 non-null int64
Question              20000 non-null object
QDescription          20000 non-null object
User                  19953 non-null object
Reputation Score      20000 non-null float64
Gold Badge Count      20000 non-null float64
Silver Badge Count    20000 non-null float64
Bronze Badge Count    20000 non-null float64
Tags                  20000 non-null object
dtypes: float64(5), int64(3), object(4)
memory usage: 2.0+ MB
```

Now our data is cleaned of missing values

We are going to ignore the missing values for user, which we are not going to use for any analysis or modelling

Now let's check for Duplicate Value

In [41]:
```
df[df.duplicated(['Question Id'], keep=False)]
```

	Question Id	Votes	Answer Count	Views	Question	QDescription	
4500	49263926	-4	2.0	21	Print Syntax Incorrect	What is wrong with the syntax here?\n\nprint '...	superloopnet
4949	49251226	0	0.0	24	Python inside MS Sql	How can i get set index value as a column insi...	
4950	49251226	0	0.0	24	Python inside MS Sql	How can i get set index value as a column insi...	
8999	49157396	1	0.0	32	Commit multiple inserts with MySQLdb on Python	I did a profiling of my python program and not...	flyingdutc
9000	49157396	1	0.0	32	Commit multiple inserts with MySQLdb on Python	I did a profiling of my python program and not...	flyingdutc
9099	49155388	0	1.0	24	Failed to Import External Dependency in Spark	I have a python script which is dependent on a...	Mav

In [42]:
```
print(len(df[df.duplicated(['Question Id'], keep=False)]))
```

```
60
```

We have 60 duplicate values

In [43]:
```
print(len(df[df.duplicated(['Question Id'], keep='first')]))
```

```
30
```

And we can remove 30 of them

In [44]:
```
a = []
```

In [45]:
```
a = df.index[df.duplicated(['Question Id'], keep='first')]
```

In [46]:
```
a
```

Out[46]:
```
Int64Index([ 4150,  4400,  4500,  4950,  9000,  9100,  9550, 10700, 10900,
            11000, 11150, 11250, 13100, 13500, 13700, 13850, 13950, 14150,
            15200, 15250, 15500, 15650, 15651, 15900, 16000, 16500, 17050,
            17250, 19450, 19700],
           dtype='int64')
```

In [47]:
```
len(a)
```

Out[47]:
```
30
```

In [48]:
```
df.drop(df.index[a[:]],inplace=True)
```

In [49]:
```
df.index[df.duplicated(['Question Id'], keep='first')]
```

Out[49]:
```
Int64Index([], dtype='int64')
```

In [50]:
```
print(len(df[df.duplicated(['Question Id'], keep='first')]))
```

```
0
```

In [51]:
```
df.info()
```

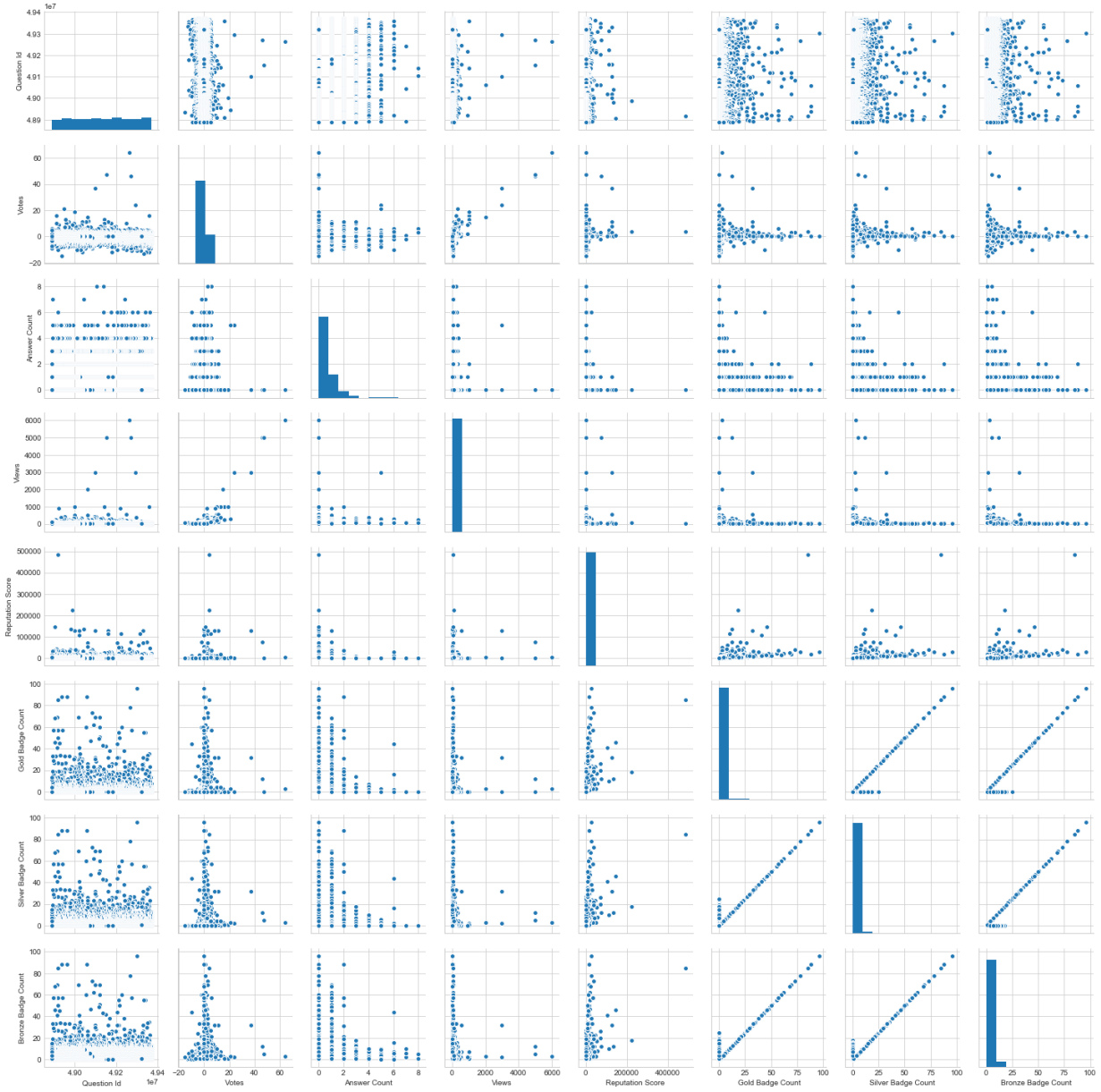
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19970 entries, 0 to 19999
Data columns (total 12 columns):
Question Id           19970 non-null int64
Votes                 19970 non-null int64
Answer Count          19970 non-null float64
Views                 19970 non-null int64
Question              19970 non-null object
QDescription          19970 non-null object
User                  19923 non-null object
Reputation Score      19970 non-null float64
Gold Badge Count      19970 non-null float64
Silver Badge Count    19970 non-null float64
Bronze Badge Count    19970 non-null float64
Tags                  19970 non-null object
dtypes: float64(5), int64(3), object(4)
memory usage: 2.0+ MB
```

Now we have removed the duplicate values

Exploratory Data Analysis to analyze data and remove outliers

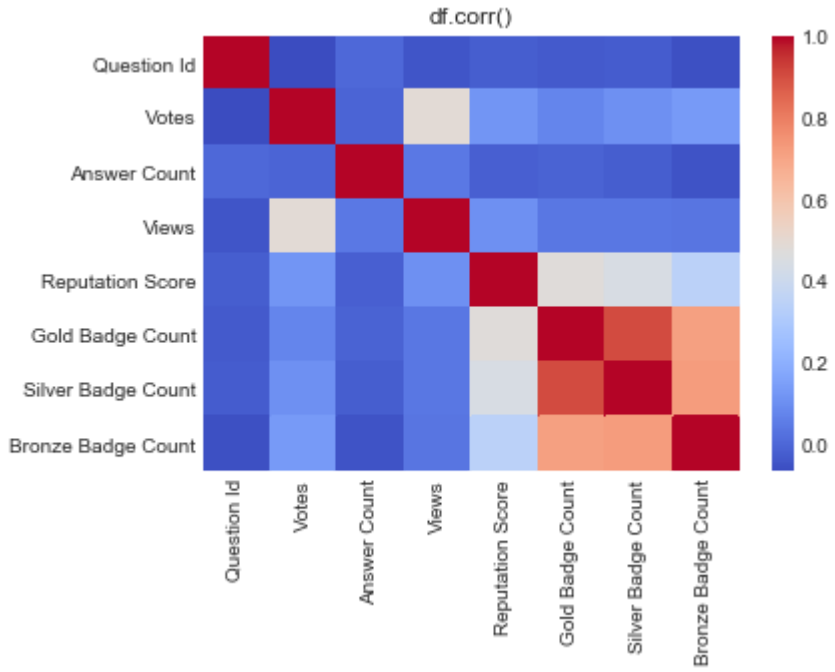
In [52]:
```
sns.pairplot(df[1:])
```

Out[52]: <seaborn.axisgrid.PairGrid at 0x203b4e84550>



In [53]:
```
#Let's analyze correlation
sns.heatmap(df.corr(),cmap='coolwarm')
plt.title('df.corr()')
```

Out[53]: Text(0.5,1,'df.corr()')



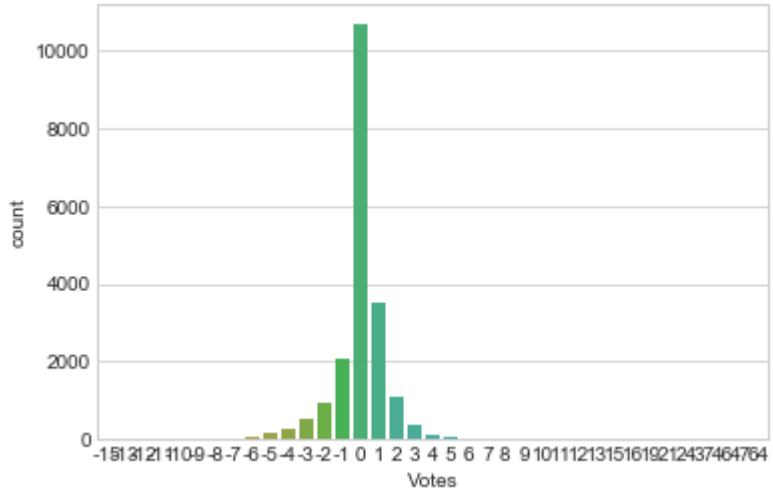
By analyisis,

1) votes are correlated with views 2) Reputation Score is Correlated with the three badges 3) Each badge is highly correlated with other and vice versa

Let's Analyze Individual Variables

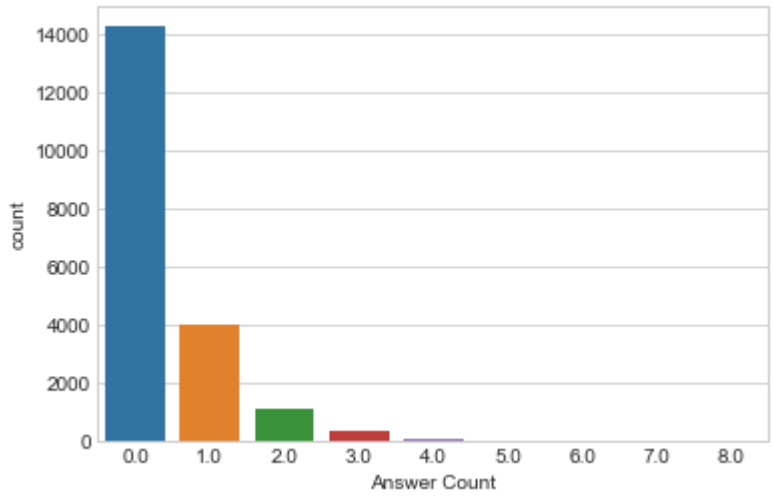
In [54]:
```
#Analyze Votes
sns.countplot(x='Votes',data=df)
```

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x203bc830438>



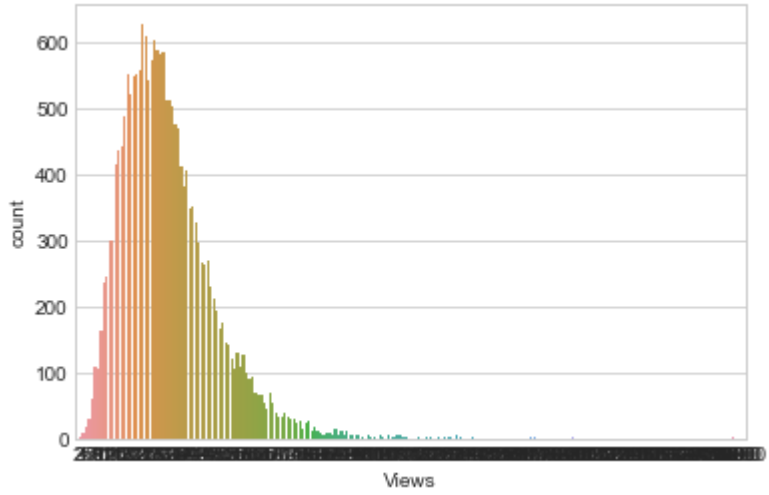
In [55]:
```
#Analyze Answer Count
sns.countplot(x='Answer Count',data=df)
```

Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x203bd31c470>



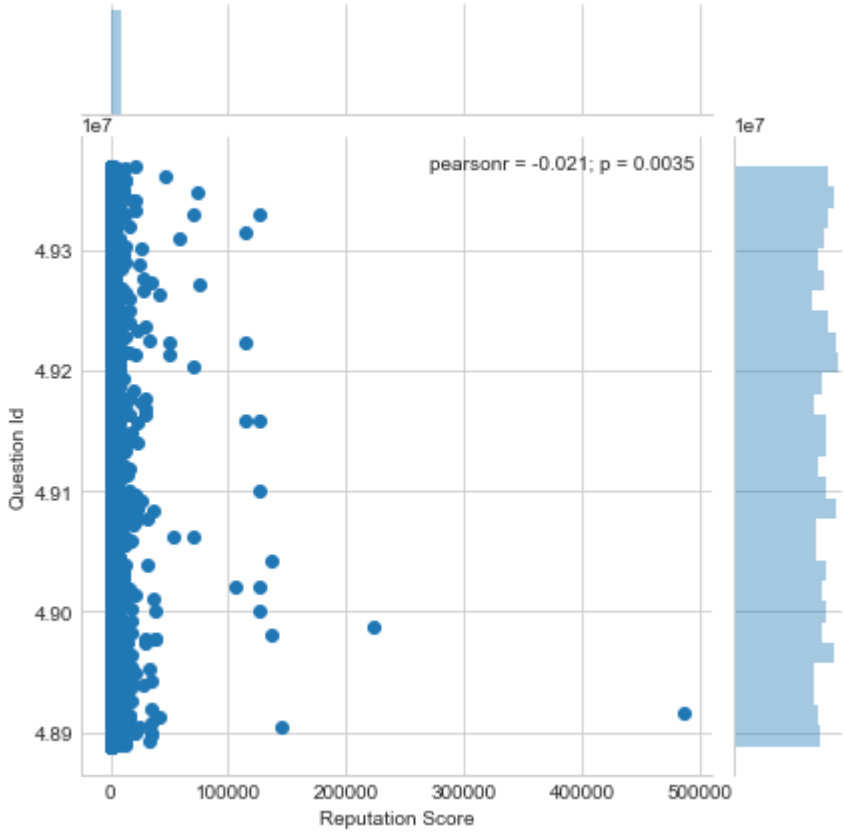
In [56]:
```
#Analyze Views
sns.countplot(x='Views',data=df)
```

Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x203bd321940>



In [57]:
```
#Exploring Reputation Score
sns.jointplot(x='Reputation Score',y='Question Id',data=df)
```

Out[57]: <seaborn.axisgrid.JointGrid at 0x203bd8900f0>



We can see Reputation Score has a outlier

In [58]:
```
df[df['Reputation Score'] > 250000]
```

Out[58]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	I
18868	48916579	4	0.0	34	Moving Collections between axes	While playing with ImportanceOfBeingErnest's c...	unutbu	486000.0	

In [59]:
```
df = df[df['Question Id'] != 48916579]
```

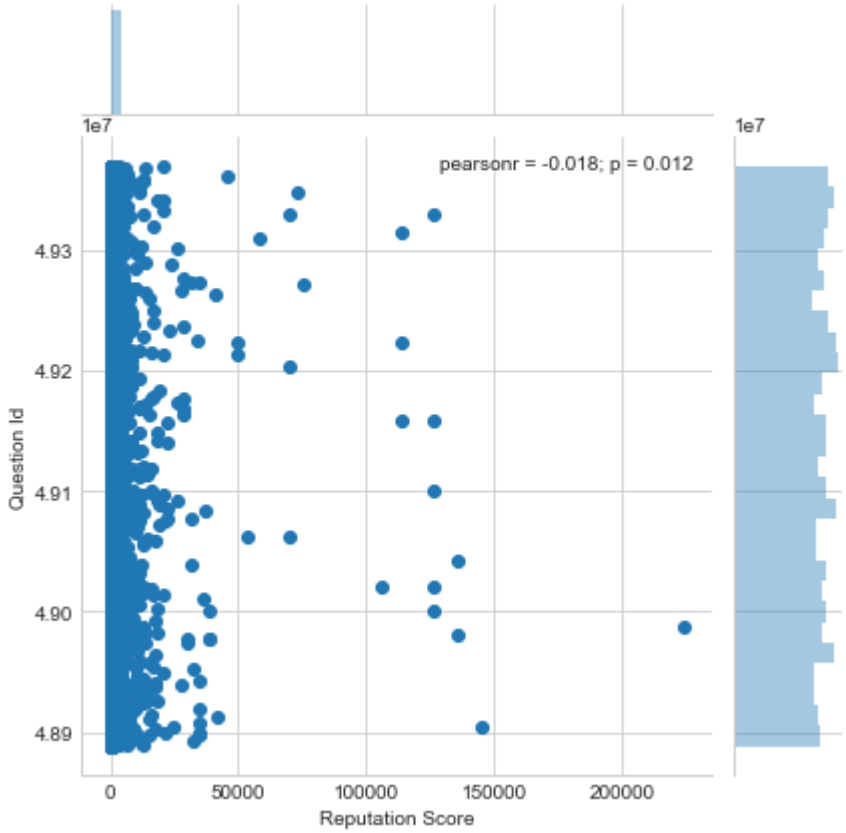
In [60]:
```
df[df['Reputation Score'] > 150000]
```

Out[60]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gol Badg Cou
15999	48988038	4	0.0	85	Find boolean mask by pattern	I have array:\n\narr = np.array([1,2,3,2,3,4,3...	jezrael	224000.0	18

In [61]:
```
sns.jointplot(x='Reputation Score',y='Question Id',data=df)
```

Out[61]: <seaborn.axisgrid.JointGrid at 0x203bd68dc50>



In [108]:
```
#df[df['Reputation Score'] > 150000]
```

In [62]:
```
df.head(1)
```

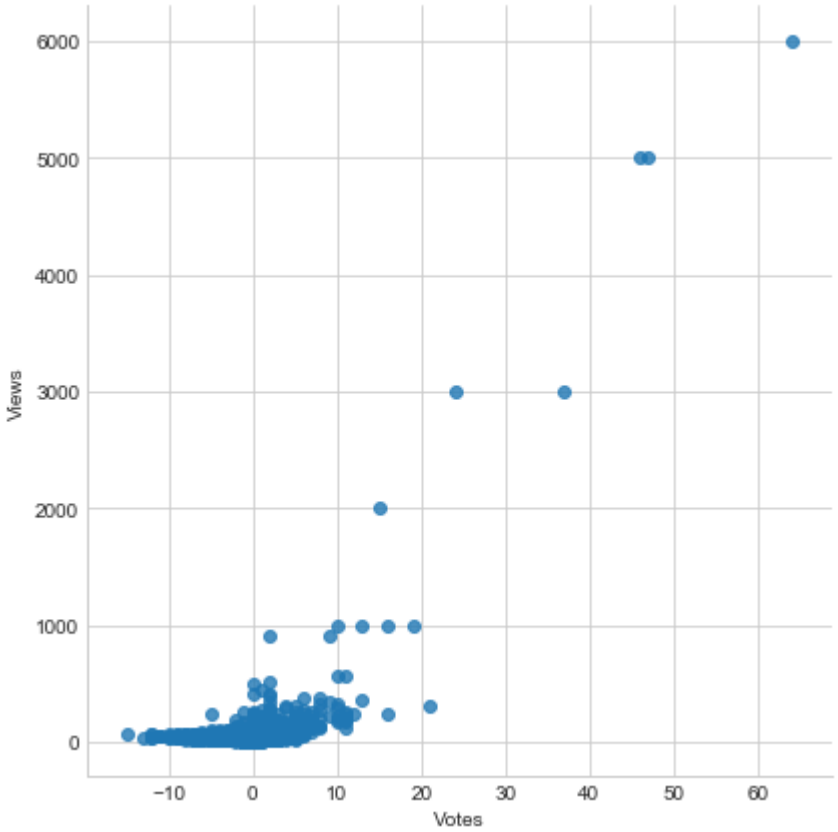
Out[62]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count	Silver Badge Count
0	49369882	0	0.0	3	Tag items with color	I have two squares in my PyOpenGL window insid...	BjkOcean	29.0	0.0	0.0

We have removed the outlier for the reputation score

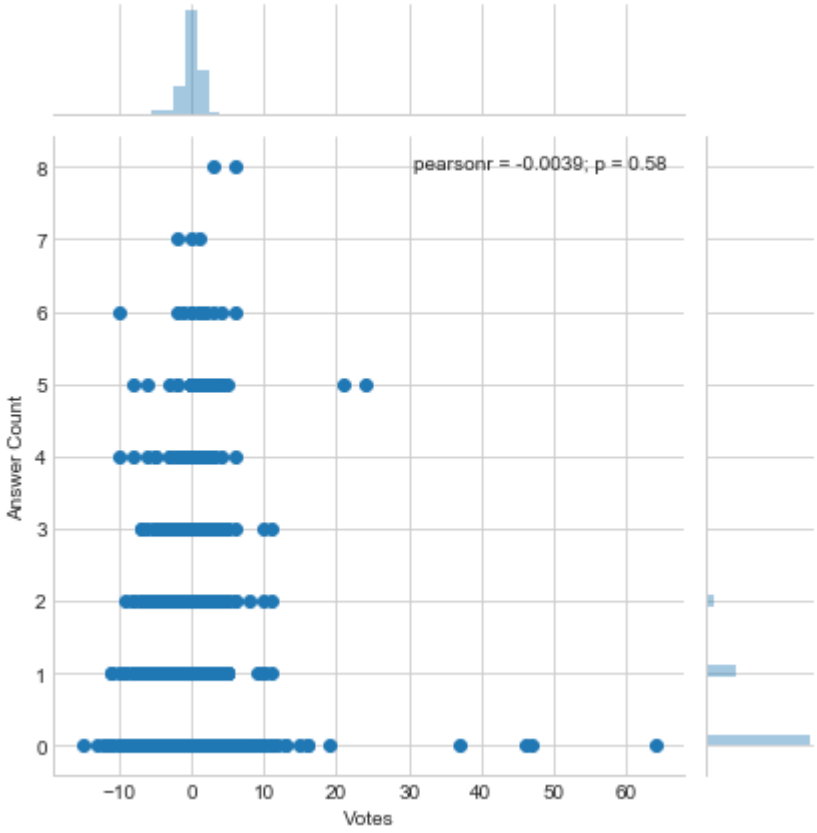
In [63]:
```
#As votes increases, views also increase
sns.set_style('whitegrid')
sns.lmplot('Votes','Views',data=df,
           palette='coolwarm',size=6,aspect=1,fit_reg=False)
```

Out[63]: <seaborn.axisgrid.FacetGrid at 0x203bea35320>



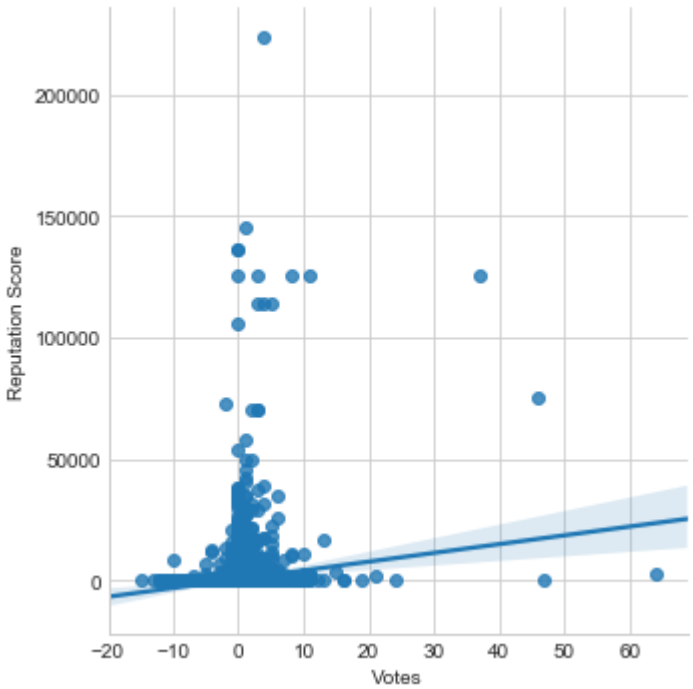
In [84]:
```
sns.jointplot(x='Votes',y='Answer Count',data=df)
```

Out[84]: <seaborn.axisgrid.JointGrid at 0x2d71c737ef0>



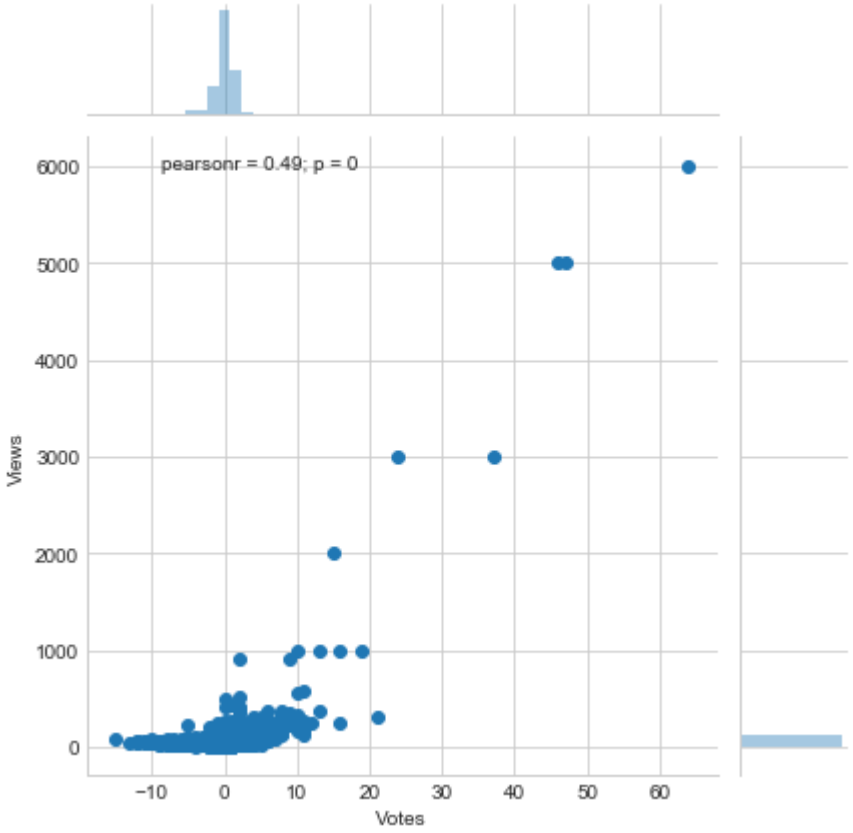
In [78]:
```
#Scatterplot of Votes vs Reputation Score
sns.lmplot(x='Votes',y='Reputation Score',data=df,palette='coolwarm')
```

Out[78]: <seaborn.axisgrid.FacetGrid at 0x203bee52c88>



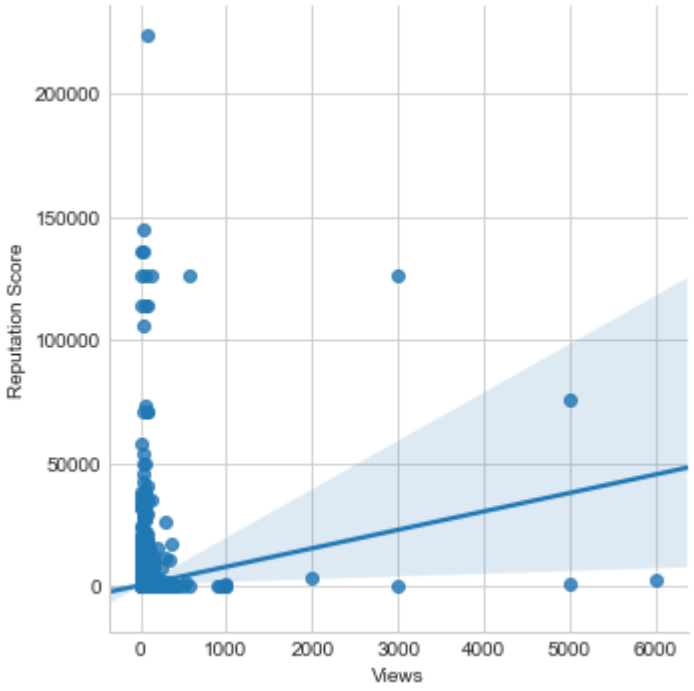
In [79]:
```
#Scatterplot of Votes vs Views
sns.jointplot(x='Votes',y='Views',data=df)
```

Out[79]: <seaborn.axisgrid.JointGrid at 0x203beecffd0>



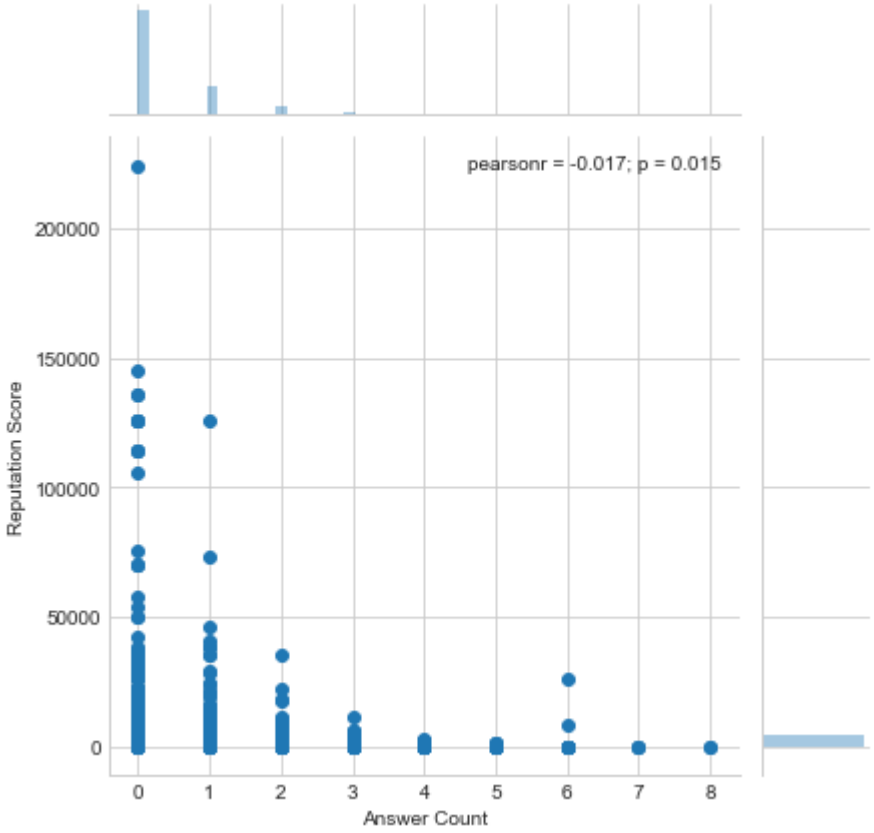
In [80]:
```
#Scatterplot of Views vs Reputation Score
sns.lmplot(x='Views',y='Reputation Score',data=df,palette='coolwarm')
```

Out[80]: <seaborn.axisgrid.FacetGrid at 0x203bf5aa630>

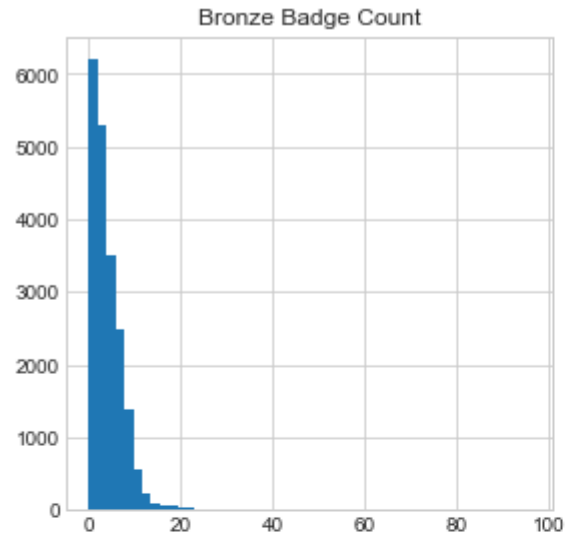


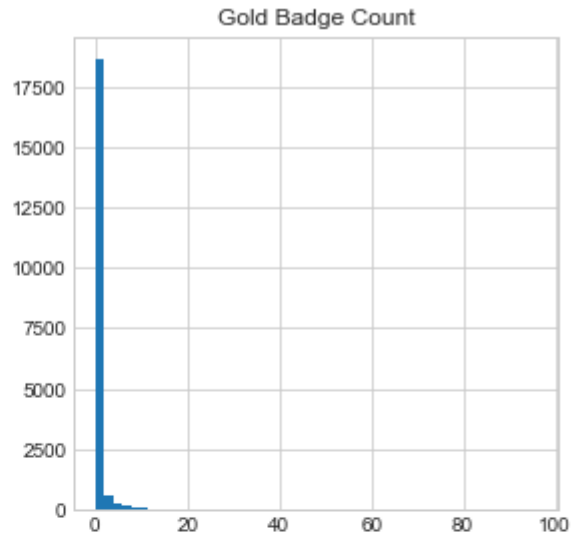
In [81]:
```
#Scatter plot of Answer Count vs Reputation Score
sns.jointplot(x='Answer Count',y='Reputation Score',data=df)
```

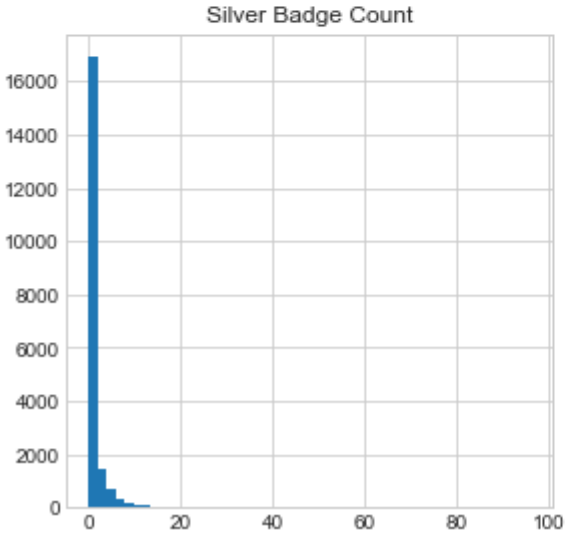
Out[81]: <seaborn.axisgrid.JointGrid at 0x203bf625390>



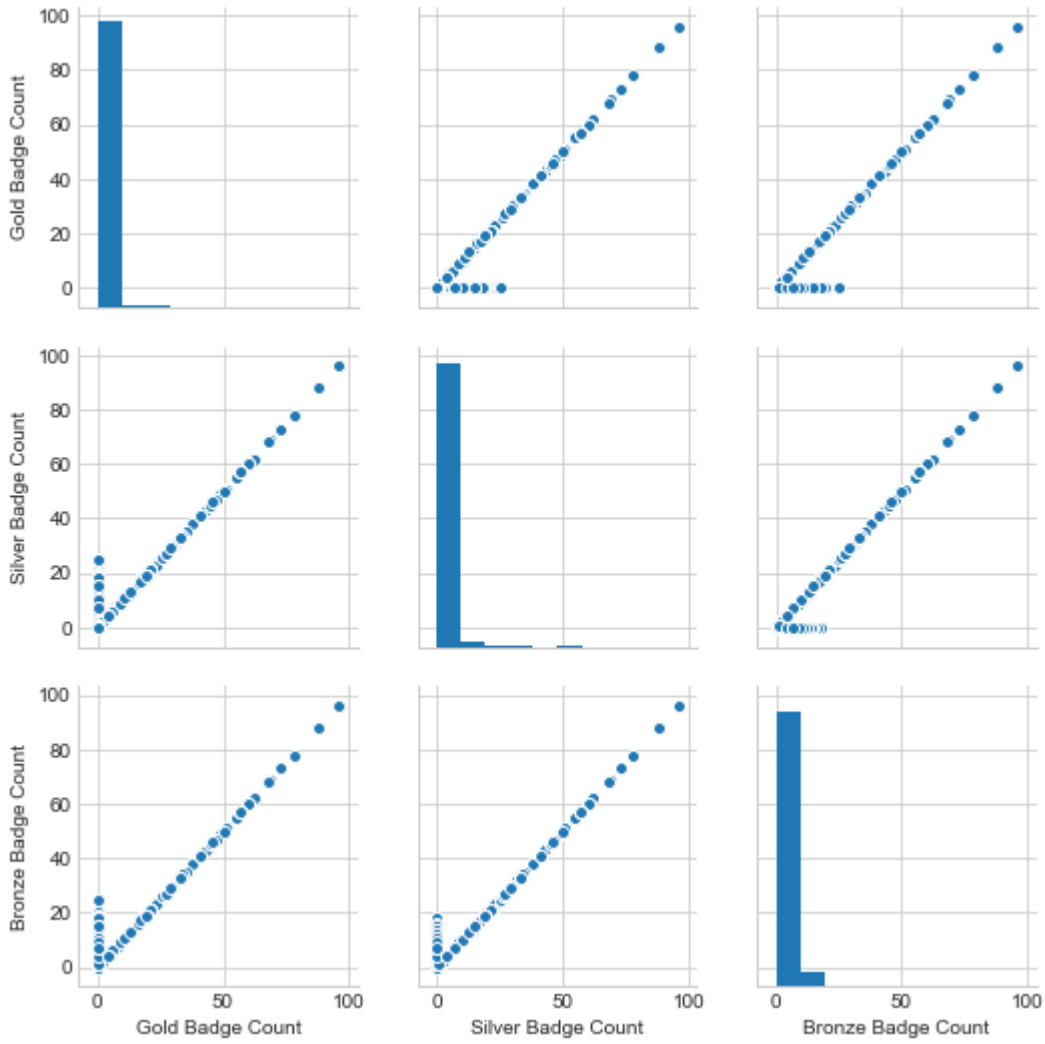
In [82]:
```
df[['Gold Badge Count', 'Silver Badge Count','Bronze Badge Count']]\
.hist(figsize=(10, 10), bins=50);
```







In [83]:
```
sns.pairplot(data=df, x_vars=['Gold Badge Count', 'Silver Badge Count','Bronze Ba
             y_vars=['Gold Badge Count', 'Silver Badge Count','Bronze Badge Count
```



Through this exploration, we are able to identify relationship between different numeric variables

Exploratory Analysis on Text Columns

In [84]:
```
df['Question'].describe()
```

Out[84]:
```
count                                19969
unique                               19969
top       numpy cumsum( ) not working?
freq                                     1
Name: Question, dtype: object
```

In [85]:
```
df['QDescription'].describe()
```

Out[85]:
```
count                                                19969
unique                                               19947
top       I have in my test module:\n\nimport pytest\nfr...
freq                                                     2
Name: QDescription, dtype: object
```

In [86]:
```
df['QDescription length'] = df['QDescription'].apply(len)
df.head()
```

Out[86]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count
0	49369882	0	0.0	3	Tag items with color	I have two squares in my PyOpenGL window insid...	BjkOcean	29.0	0.0
1	49369867	0	0.0	5	Firebase dump json data	I'm no back-end developer. So perspective is a...	user9132502	52.0	0.0
2	49369855	-5	0.0	14	Specific type of webscraping [on hold]	How can I get python to check a specific line ...	Ronprogramming	4.0	0.0
3	49369846	0	0.0	3	Notification on android via Python using Pushs...	I am using "Pushsafer" to notify on my smartph...	Sanket	1.0	0.0
4	49369842	0	0.0	14	How to get the exact count of people by face d...	I am working on getting the total count of peo...	Mueez Siraj	101.0	0.0

In [96]:
```
#df.drop('length',axis=1,inplace=True)
```

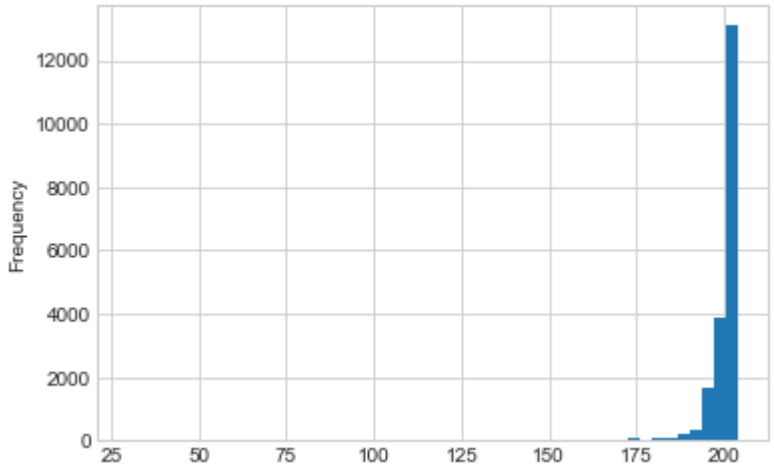
In [87]:
```
df.head()
```

Out[87]:

	Question Id	Votes	Answer Count	Views	Question	QDescription	User	Reputation Score	Gold Badge Count
0	49369882	0	0.0	3	Tag items with color	I have two squares in my PyOpenGL window insid...	BjkOcean	29.0	0.0
1	49369867	0	0.0	5	Firebase dump json data	I'm no back-end developer. So perspective is a...	user9132502	52.0	0.0
2	49369855	-5	0.0	14	Specific type of webscraping [on hold]	How can I get python to check a specific line ...	Ronprogramming	4.0	0.0
3	49369846	0	0.0	3	Notification on android via Python using Pushs...	I am using "Pushsafer" to notify on my smartph...	Sanket	1.0	0.0
4	49369842	0	0.0	14	How to get the exact count of people by face d...	I am working on getting the total count of peo...	Mueez Siraj	101.0	0.0

In [88]:
```
df['QDescription length'].plot(bins=50, kind='hist')
```

Out[88]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x203c0984470>
```



In [89]:
```
df['QDescription length'].describe()
```

Out[89]:
```
count    19969.000000
mean       199.364265
std         11.207627
min         30.000000
25%        199.000000
50%        202.000000
75%        203.000000
max        204.000000
Name: QDescription length, dtype: float64
```

Explore tags

In [90]:
```
print(df['Tags'].unique())
```

```
['python pyqt4 pyopengl' 'android python firebase'
 'python python-3.x web-scraping' ...,
 'python tensorflow deep-learning tensorboard' 'python numpy random seed'
 'python unity3d controls']
```

In [96]:
```
print(df['Tags'].value_counts())
```

```
python                                  1207
python pandas                            579
python python-3.x                        473
python django                            259
python pandas dataframe                  210
python tensorflow                        184
python numpy                             143
python python-2.7                        135
python regex                             129
python matplotlib                        121
python tkinter                           104
python list                               84
python dictionary                         58
python pandas numpy                       55
python json                               55
python opencv                             54
python beautifulsoup                      50
python python-3.x pandas                  50
python arrays numpy                       50
python tensorflow keras                   50
python flask                              48
python selenium                           45
python pandas pandas-groupby              45
python csv                                42
python python-3.x tkinter                 40
python django django-rest-framework       39
python scikit-learn                       37
python sqlalchemy
```

```
     35
python python-3.x python-2.7
     34
python string
     32

   ...
python matplotlib converter plotly heatmap
      1
python conditional branch
      1
python xml python-3.x xml-parsing
      1
python optimization black-box
      1
python pandas dataframe replace data-manipulation
      1
javascript python selenium web-scraping
      1
python tkinter rotation tkinter-canvas
      1
python scripting gtk
      1
python python-3.x list file dictionary
      1
python django forms post model
      1
python function tkinter parameters
      1
python selenium email extract
      1
python image scrapy pipeline scrapy-spider
      1
python serial-port
      1
python ironpython language-design dynamic-language-runtime dynamic-languages
      1
python rectangles
      1
python algorithm quicksort
      1
python dry
      1
python python-3.x sorting numpy group-by
      1
python jinja2 jupyter-notebook jupyter
      1
python nlp deep-learning keras
      1
python server xampp cv2
      1
python numpy matrix-multiplication tensor numpy-ndarray
      1
python sockets http http-headers
      1
python paramiko distributed-system rpyc
      1
```

```
python hierarchical-clustering em
      1
python ironpython python.net
      1
python fortran ctypes gfortran
      1
python network-programming sys python-sockets python-os
      1
python tkinter tkinter-layout
      1
Name: Tags, Length: 11686, dtype: int64
```

Wordclouds for Text Data

In [102]:
```
from subprocess import check_output
from wordcloud import WordCloud, STOPWORDS

#mpl.rcParams['figure.figsize']=(8.0,6.0)    #(6.0,4.0)
mpl.rcParams['font.size']=12                #10
mpl.rcParams['savefig.dpi']=100             #72
mpl.rcParams['figure.subplot.bottom']=.1


stopwords = set(STOPWORDS)
#data = pd.read_csv("../input/most_backed.csv")

wordcloud = WordCloud(
                          background_color='white',
                          stopwords=stopwords,
                          max_words=200,
                          max_font_size=40,
                          random_state=42
                         ).generate(str(df['Question']))

print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
#fig.savefig("word1.png", dpi=900)
```

```
<wordcloud.wordcloud.WordCloud object at 0x00000203C093B518>
```

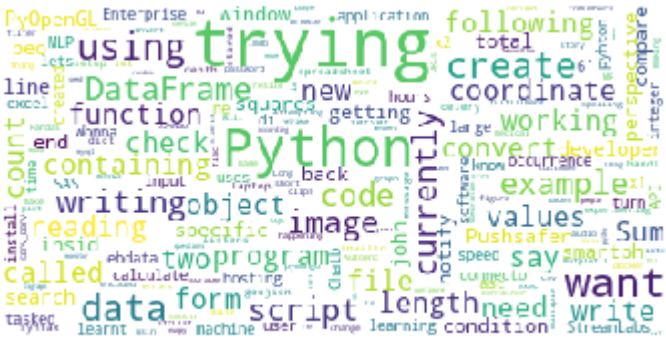
In [103]:
```
#mpl.rcParams['figure.figsize']=(8.0,6.0)    #(6.0,4.0)
mpl.rcParams['font.size']=12                #10
mpl.rcParams['savefig.dpi']=200             #72
mpl.rcParams['figure.subplot.bottom']=.1


stopwords = set(STOPWORDS)
#data = pd.read_csv("../input/most_backed.csv")

wordcloud = WordCloud(
                          background_color='white',
                          stopwords=stopwords,
                          max_words=200,
                          max_font_size=40,
                          random_state=42
                         ).generate(str(df['QDescription']))

print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
#fig.savefig("word1.png", dpi=900)
```

```
<wordcloud.wordcloud.WordCloud object at 0x00000203C08E9AC8>
```



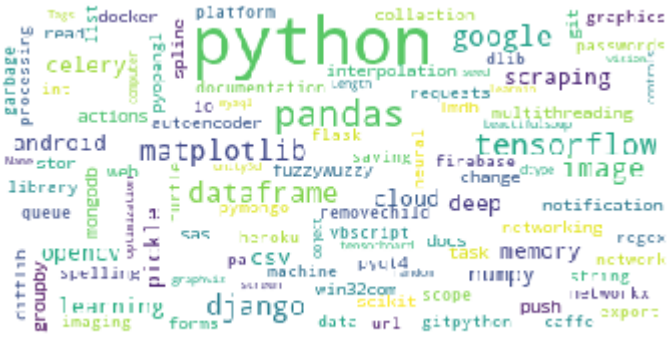
In [104]:
```
#mpl.rcParams['figure.figsize']=(8.0,6.0)    #(6.0,4.0)
mpl.rcParams['font.size']=12                #10
mpl.rcParams['savefig.dpi']=200             #72
mpl.rcParams['figure.subplot.bottom']=.1


stopwords = set(STOPWORDS)
#data = pd.read_csv("../input/most_backed.csv")

wordcloud = WordCloud(
                          background_color='white',
                          stopwords=stopwords,
                          max_words=200,
                          max_font_size=40,
                          random_state=42
                         ).generate(str(df['Tags']))

print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
#fig.savefig("word1.png", dpi=900)
```

```
<wordcloud.wordcloud.WordCloud object at 0x00000203C09F9B00>
```



Storing Data in Excel File

In [105]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19969 entries, 0 to 19999
Data columns (total 13 columns):
Question Id            19969 non-null int64
Votes                  19969 non-null int64
Answer Count           19969 non-null float64
Views                  19969 non-null int64
Question               19969 non-null object
QDescription           19969 non-null object
User                   19922 non-null object
Reputation Score       19969 non-null float64
Gold Badge Count       19969 non-null float64
Silver Badge Count     19969 non-null float64
Bronze Badge Count     19969 non-null float64
Tags                   19969 non-null object
QDescription length    19969 non-null int64
dtypes: float64(5), int64(4), object(4)
memory usage: 2.8+ MB
```

In [85]:
```
from pandas import ExcelWriter

writer = ExcelWriter('Stack_Overflow_Questions_Clean_EDA_Data.xlsx')
df.to_excel(writer,'Sheet1')
writer.save()
```

In [ ]:
```

```