DATA ANALYSIS ASSIGNMENT

Background:

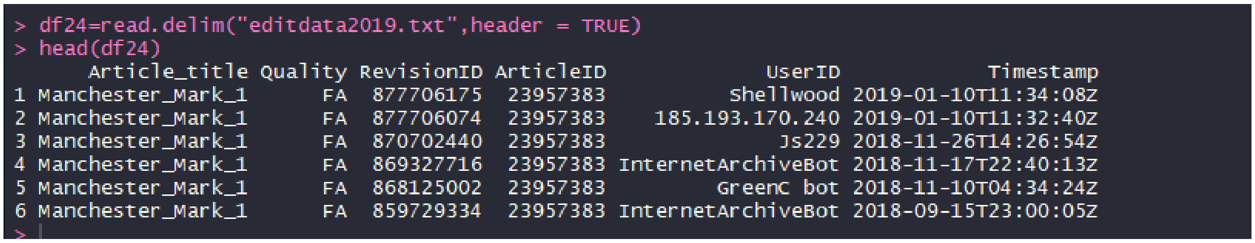
- The focus of this assignment is to compare the quality of a sampled group of articles on Wikipedia based on a few parameters
- The features available to us are:
    - Article_title: title of the article
    - Article ID: unique ID of the article, each article title will have one correspondingly
    - Revision ID: Unique ID of each edit done by users randomly generated but sequential
    - Quality: Has 4 possible values with 'C' being Bad and the rest being acceptable or good (FA,GA,A)
    - User ID: This is the feature to differentiate between which user is committing the edit
    - Timestamp: Once the edit is confirmed ,this timestamp is logged in the system corresponding to the revision ID

Method:

- First step was to conduct a comprehensive evaluation of the data with multiple parameters pitted against each other with **Quality** being the target feature
- Since the main goal is to analyze based on Box plots for 5 different scenarios, this report will focus on that.
- I downloaded the data from the dropbox link and after going through the data to get a clear idea of the shape, I imported the dataset into tableau
- For each scenario, I used the appropriate filters and context on the worksheet before exporting the data file accordingly

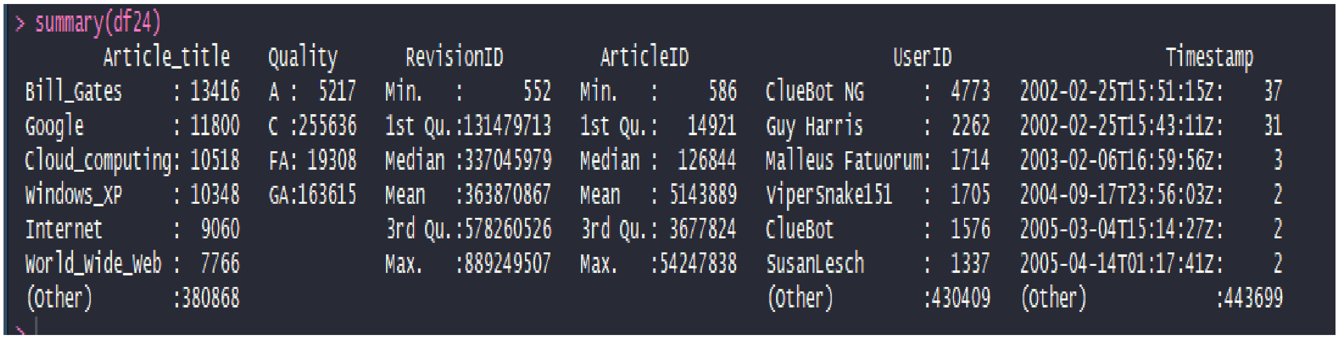
Brief Snapshots(using R):

Data Outlook:



```
> df24=read.delim("editdata2019.txt",header = TRUE)
> head(df24)
     Article_title Quality RevisionID ArticleID              UserID            Timestamp
1 Manchester_Mark_1      FA  877706175  23957383           Shellwood 2019-01-10T11:34:08Z
2 Manchester_Mark_1      FA  877706074  23957383    185.193.170.240 2019-01-10T11:32:40Z
3 Manchester_Mark_1      FA  870702440  23957383               Js229 2018-11-26T14:26:54Z
4 Manchester_Mark_1      FA  869327716  23957383 InternetArchiveBot 2018-11-17T22:40:13Z
5 Manchester_Mark_1      FA  868125002  23957383          GreenC bot 2018-11-10T04:34:24Z
6 Manchester_Mark_1      FA  859729334  23957383 InternetArchiveBot 2018-09-15T23:00:05Z
> 
```

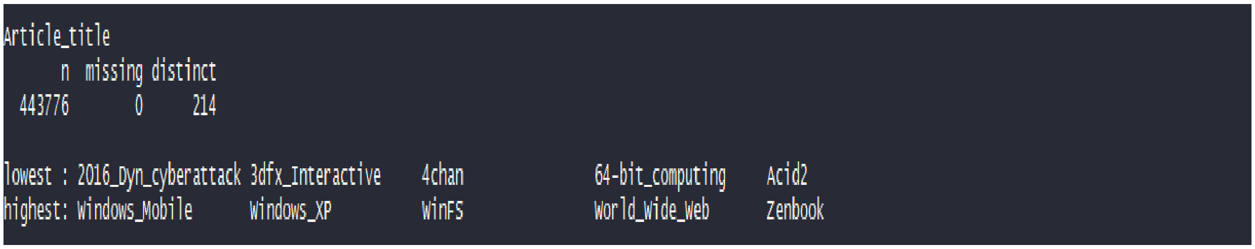
Summary:



```
> summary(df24)
        Article_title     Quality        RevisionID          ArticleID                    UserID                   Timestamp
 Bill_Gates     : 13416   A :  5217   Min.   :      552   Min.   :      586   ClueBot NG     :  4773   2002-02-25T15:51:15Z:    37
 Google         : 11800   C :255636   1st Qu.:131479713   1st Qu.:   14921   Guy Harris     :  2262   2002-02-25T15:43:11Z:    31
 Cloud_computing: 10518   FA: 19308   Median :337045979   Median :  126844   Malleus Fatuorum:  1714   2003-02-06T16:59:56Z:     3
 Windows_XP     : 10348   GA:163615   Mean   :363870867   Mean   : 5143889   ViperSnake151  :  1705   2004-09-17T23:56:03Z:     2
 Internet       :  9060               3rd Qu.:578260526   3rd Qu.: 3677824   ClueBot        :  1576   2005-03-04T15:14:27Z:     2
 World_Wide_Web :  7766               Max.   :889249507   Max.   :54247838   SusanLesch     :  1337   2005-04-14T01:17:41Z:     2
 (Other)        :380868                                                      (Other)        :430409   (Other)             :443699
```

- It's clear from the data that Bill Gates is the most edited title along with Google and Cloud_computing coming close
- Individually C level articles have the highest proportion overall in Quality, which might suggest higher number of edits -> lesser quality (possible)
- ClueBot NG , Guy Harris and Malleus Faturom are the highest contributors

- **Article_Title**

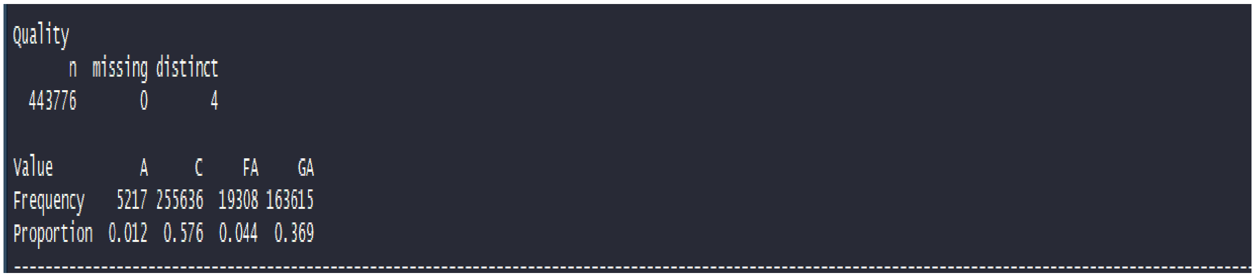


```
Article_title
       n  missing distinct
  443776        0      214

lowest : 2016_Dyn_cyberattack 3dfx_Interactive     4chan                64-bit_computing     Acid2
highest: Windows_Mobile       Windows_XP           WinFS                World_Wide_Web       Zenbook
```

- There are 214 different articles in the data

- **Quality**

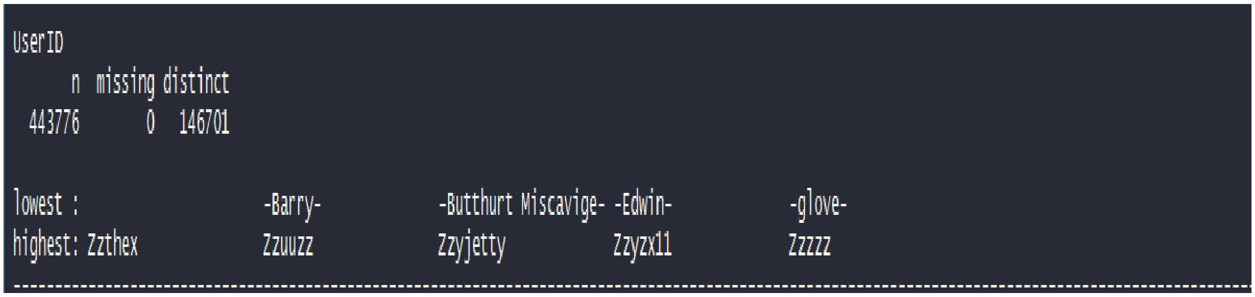


```
Quality
       n  missing distinct
  443776        0        4

Value           A      C     FA     GA
Frequency    5217 255636  19308 163615
Proportion  0.012  0.576  0.044  0.369
------------------------------------------------------------------------------------------------------------------------------------------
```

- 'C' level articles have 57.6% of the data
- 'A' level articles form around 1.2% of the data

- **User ID**



```
UserID
       n  missing distinct
  443776        0   146701

lowest :                  -Barry-          -Butthurt Miscavige- -Edwin-          -glove-
highest: Zzthex           Zzuuzz           Zzyjetty             Zzyzx11          Zzzzz
--------------------------------------------------------------------------------------------------------
```

- There are 146701 unique/distinct users who have committed edits

Next, let's move on the first scenario,

1. Total number of unique editors

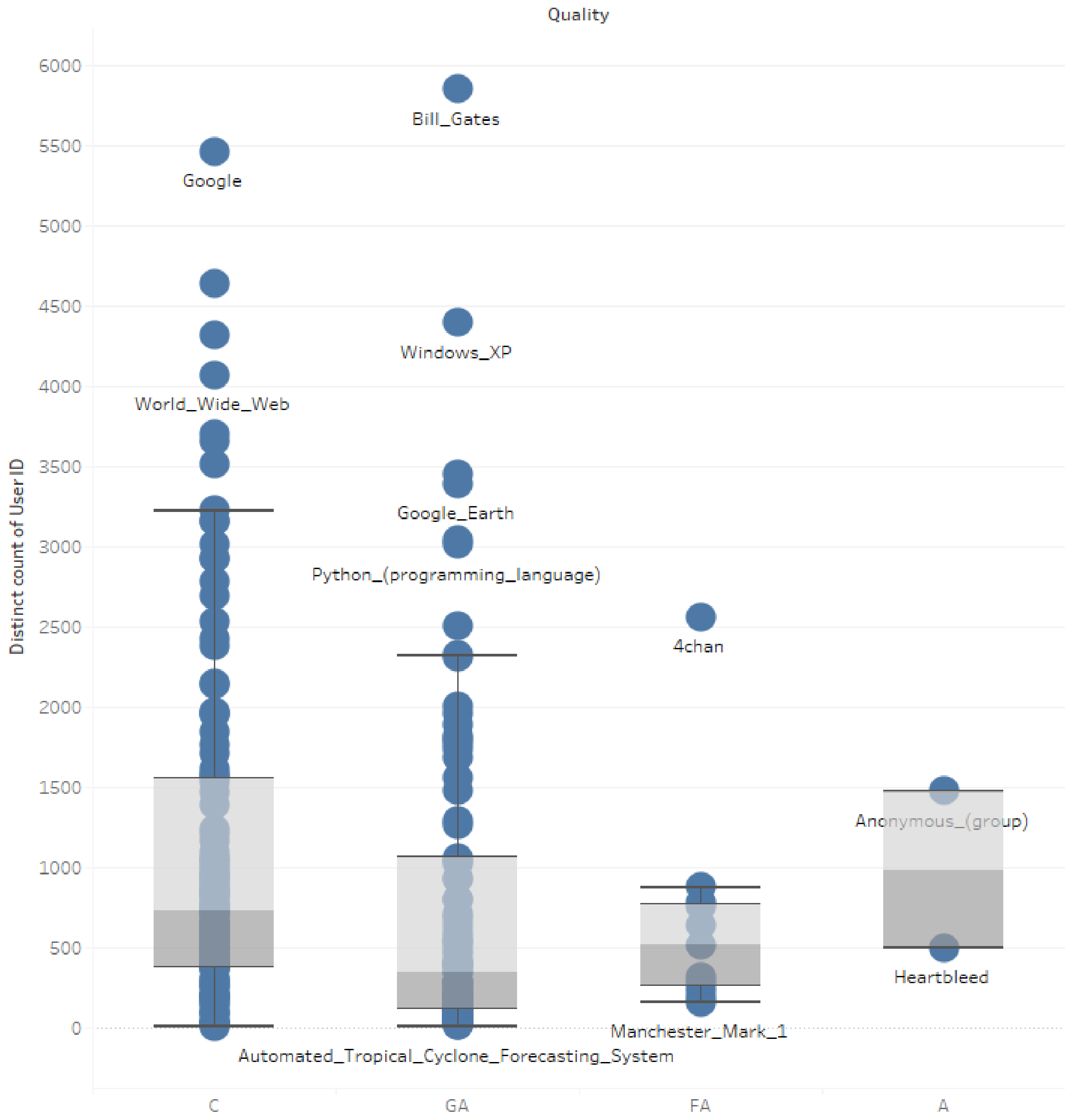
Initial Hypothesis: "Too many cooks spoil the broth" . More unique editors should mean lower standard of quality

- In Tableau, after importing the data, I created a sheet where I considered the following:
  - I wanted the labels on the graph to be the article title
  - The columns would be the Quality of the article with 'C' coming first and the rest coming in the rear
  - The rows would be distinct userID count for the articles, hence higher unique editors, the article would appear higher on the graph
  - The data after filtering, I exported as a CSV file called " 1.DistinctUserIDvsQuality"

Box Plot:

- The numbers are the number of unique editors (each article is of only 1 quality corresponding to it ,not a time series dataset where quality keeps changing)

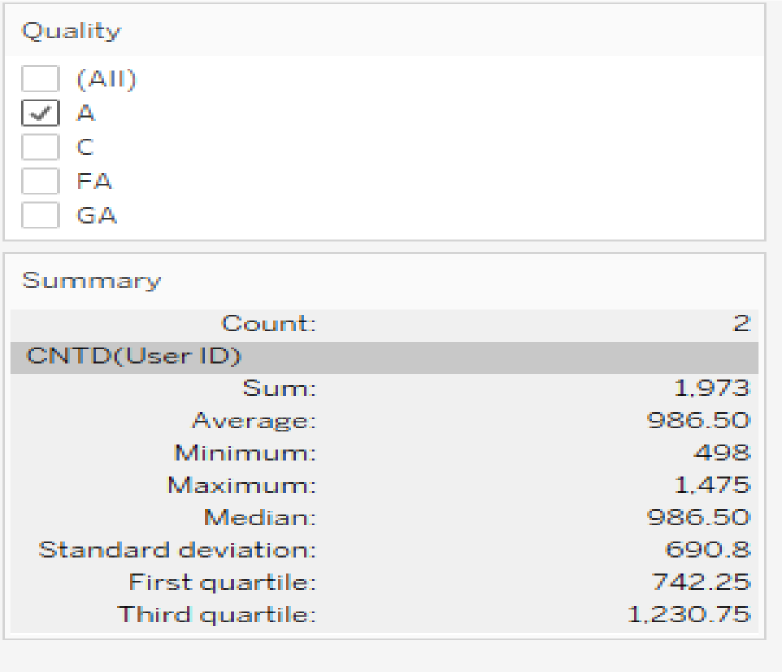
1.DistinctUserIDvsQuality



Distinct count of User ID for each Quality. The marks are labeled by Article title.

Statistics:

For Type A:



Quality

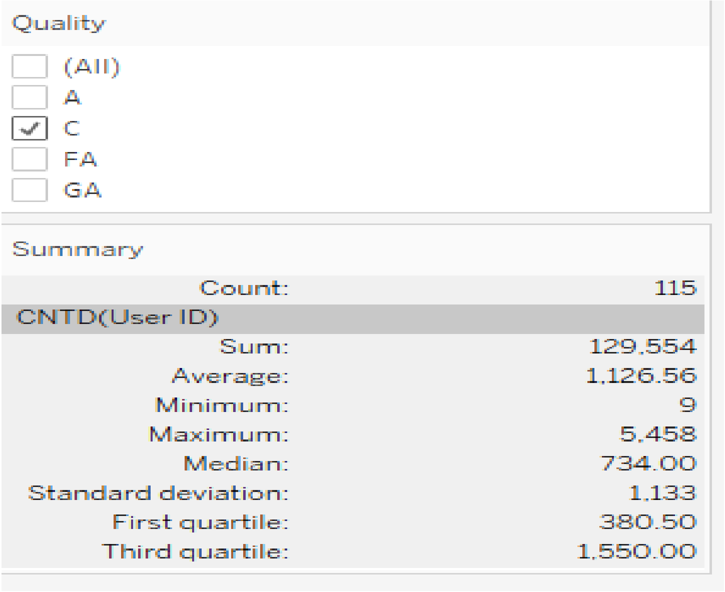
- ☐ (All)
- ☑ A
- ☐ C
- ☐ FA
- ☐ GA

Summary

Count:	2
CNTD(User ID)	
Sum:	1,973
Average:	986.50
Minimum:	498
Maximum:	1,475
Median:	986.50
Standard deviation:	690.8
First quartile:	742.25
Third quartile:	1,230.75

- There are only 2 articles, with 1973 unique editors
- The Median and average are both 986.50
- Standard Deviation : 690.8

For Type C:



Quality

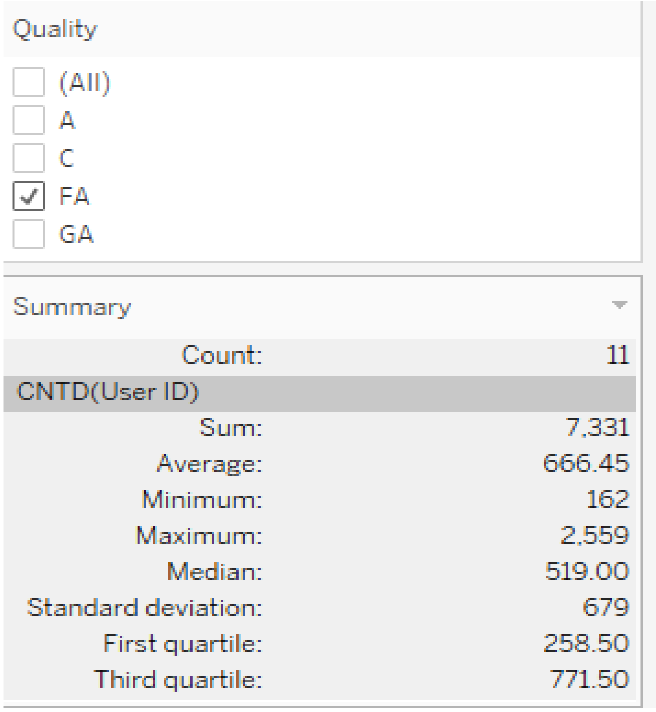
- ☐ (All)
- ☐ A
- ☑ C
- ☐ FA
- ☐ GA

Summary

Count:	115
CNTD(User ID)	
Sum:	129,554
Average:	1,126.56
Minimum:	9
Maximum:	5,458
Median:	734.00
Standard deviation:	1,133
First quartile:	380.50
Third quartile:	1,550.00

- Here there are 115 articles , with Average being 1126.56 and Median being 734
- The Standard deviation is 1.133

For Type FA:



Quality

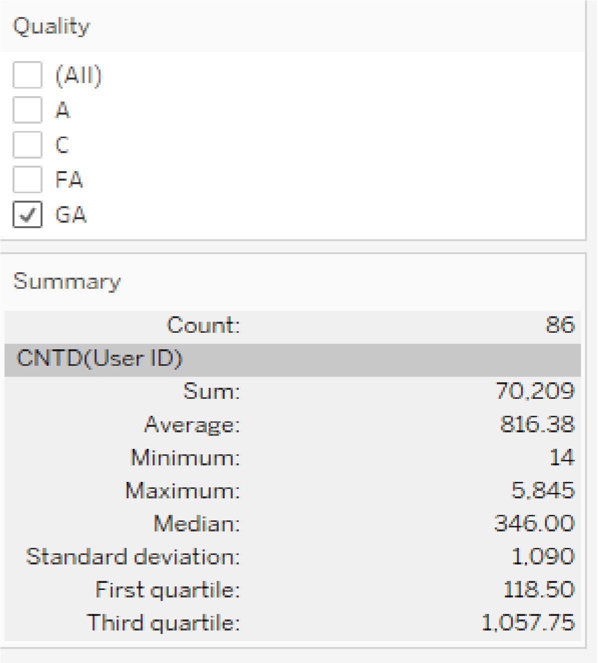
- □ (All)
- □ A
- □ C
- ☑ FA
- □ GA

Summary

Count:	11
CNTD(User ID)	
Sum:	7,331
Average:	666.45
Minimum:	162
Maximum:	2,559
Median:	519.00
Standard deviation:	679
First quartile:	258.50
Third quartile:	771.50

- The average is 666.45 and Standard deviation is 679. The median number of unique users is 519
- Only 11 articles here being considered

For Type GA:



Quality

- □ (All)
- □ A
- □ C
- □ FA
- ☑ GA

Summary

Count:	86
CNTD(User ID)	
Sum:	70,209
Average:	816.38
Minimum:	14
Maximum:	5,845
Median:	346.00
Standard deviation:	1,090
First quartile:	118.50
Third quartile:	1,057.75

- For Type GA there are 86 cases and the median is 346 with the average being 816.38
- Standard Deviation is 1,090

Conclusion:

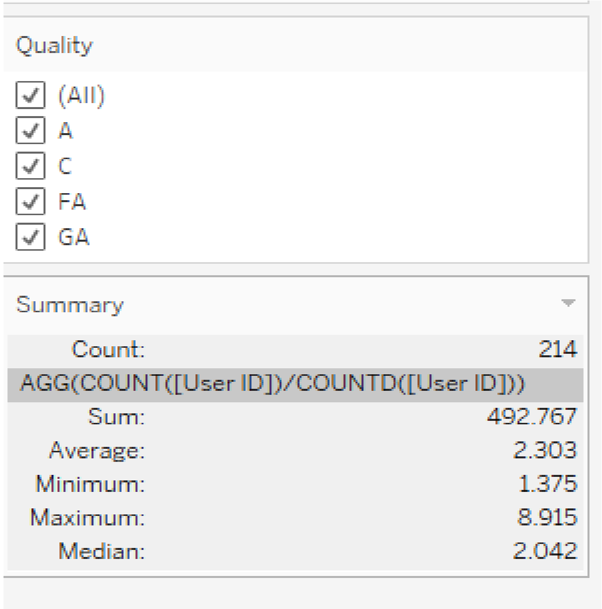
- **Type C has a higher Average number of Unique users as well as higher median (not considering type A as it has only 2 cases)**
- **Outliers :**
    - **In type C , some of the articles such as Google,Cloud_computing have higher number of unique users, this doesn't contrast with my hypothesis . Some of the cases have lower numbers like ACM_SIGHPC , but this started in 2017, so it's more recent. Adoption_Software_Implementation has 24 unique editors, here the article may not have enough clout to have too many editors**
    - **In type_GA , we see Bill Gates ,Windows XP ,Google Earth,Python and twitter having high unique editors. These articles have more important topics hence a lot more edits**
    - **In Type FA, 4chan has a lot of unique edits, this is a more controversial topic and will have a lot of sections consistently being updated.**
    - **Type A has only 2 articles so no real outliers**

2. Number of edits per editors

Hypothesis: Higher edit average would mean better quality as usually untrusted/unverified editors wouldn't edit multiple times.

- Once again, the column here is quality
- For Rows, it gets tricky, using an aggregate function which looks like
    - COUNT([User ID])/COUNTD([User ID])
- The data filtered, has been exported as a CSV file called "2.Number of Edits per Editor"

OVERALL:



Quality

☑ (All)
☑ A
☑ C
☑ FA
☑ GA

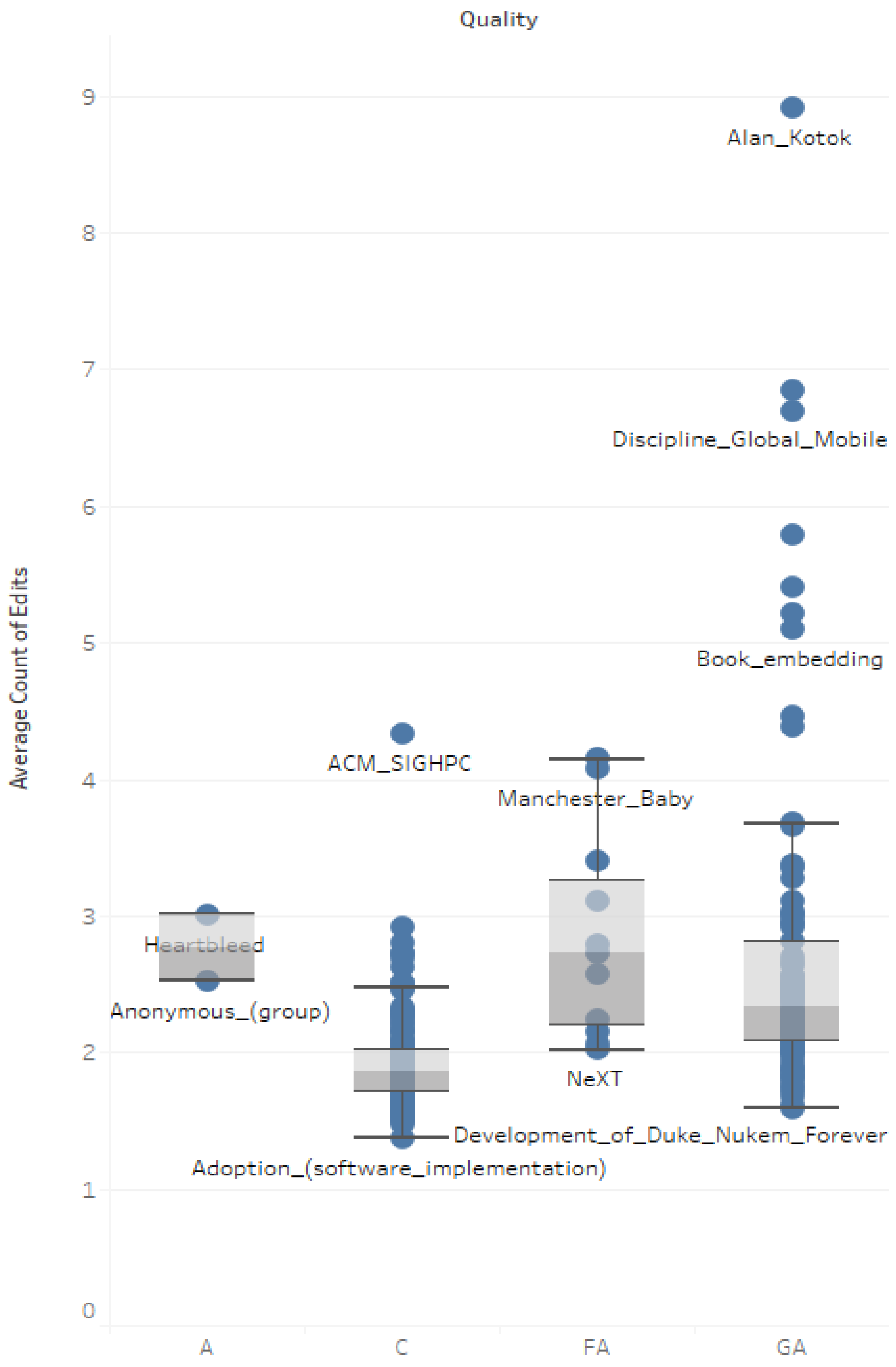
Summary

Count:	214
AGG(COUNT([User ID])/COUNTD([User ID]))	
Sum:	492.767
Average:	2.303
Minimum:	1.375
Maximum:	8.915
Median:	2.042

- The average is 2.303 for articles
- Median is 2.042

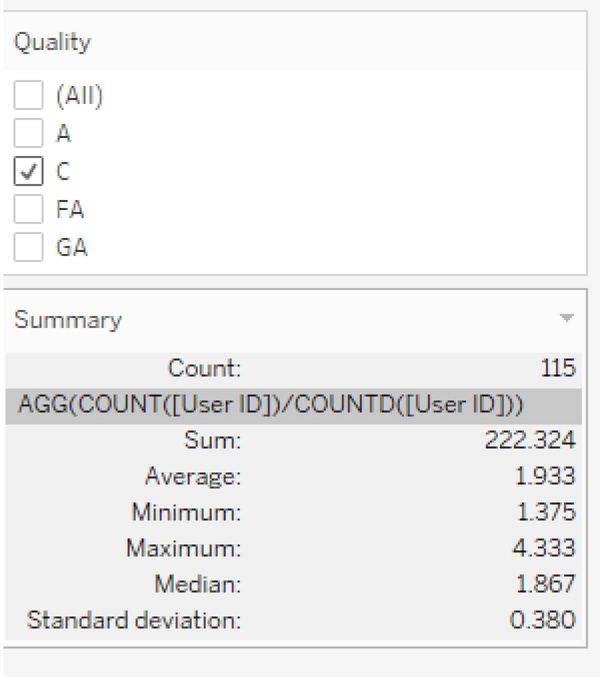
BOX PLOT

2.NumberofEditsPerEditor (2)



COUNT([User ID])/COUNTD([User ID]) for each Quality. The marks are labeled by Article title. Details are shown for Article title. The view is filtered on Article title and Quality. The Article title filter keeps 214 of 214 members. The Quality filter keeps A, C, FA and GA.

- Type C:



Quality

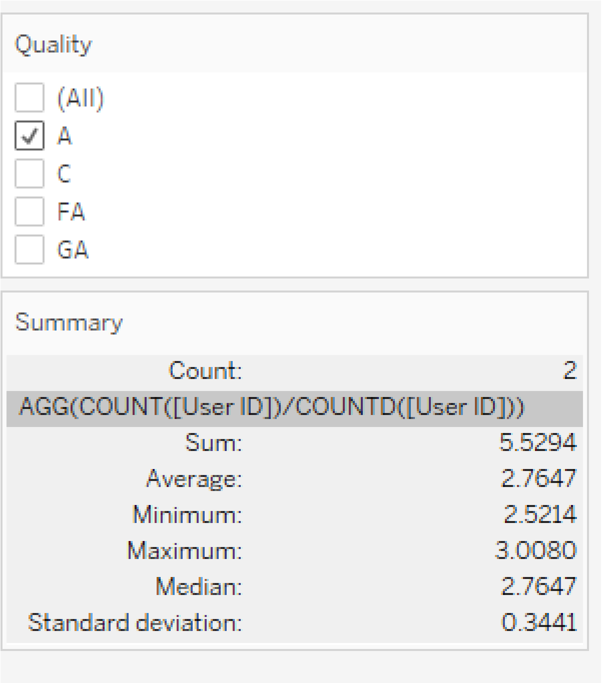
☐ (All)
☐ A
☑ C
☐ FA
☐ GA

Summary

Count:	115
AGG(COUNT([User ID])/COUNTD([User ID]))	
Sum:	222.324
Average:	1.933
Minimum:	1.375
Maximum:	4.333
Median:	1.867
Standard deviation:	0.380

- Here the average is 1.9333, while the median is 1.867
- Standard deviation is 0.38

- Type A



Quality

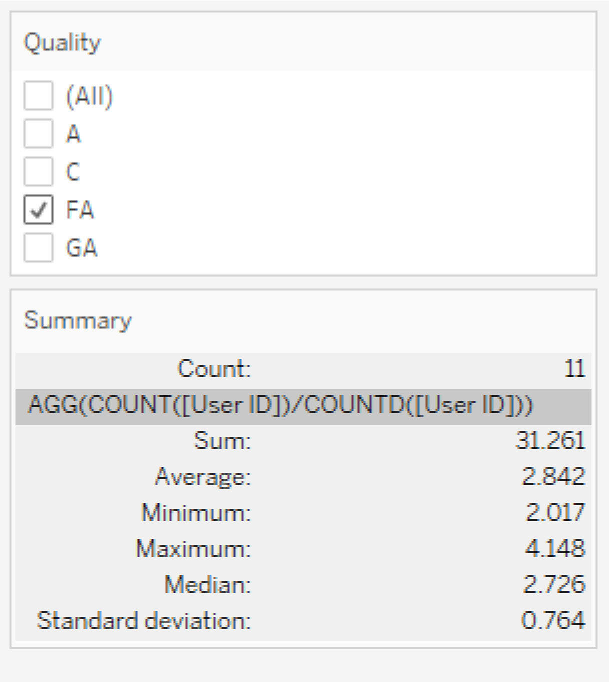
☐ (All)
☑ A
☐ C
☐ FA
☐ GA

Summary

Count:	2
AGG(COUNT([User ID])/COUNTD([User ID]))	
Sum:	5.5294
Average:	2.7647
Minimum:	2.5214
Maximum:	3.0080
Median:	2.7647
Standard deviation:	0.3441

- Average is 2.764 and Median is the same as there are only 2 cases

- Type FA



Quality

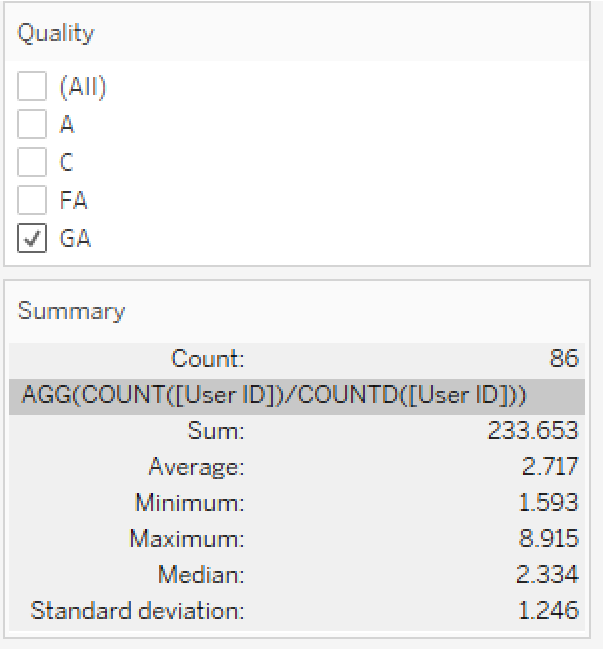
☐ (All)
☐ A
☐ C
☑ FA
☐ GA

Summary

Count:	11
AGG(COUNT([User ID])/COUNTD([User ID]))	
Sum:	31.261
Average:	2.842
Minimum:	2.017
Maximum:	4.148
Median:	2.726
Standard deviation:	0.764

- Here average is 2.8 , median is 2.726 while standard deviation is 0.764
- There are 11 cases here

- Type GA



Quality

☐ (All)
☐ A
☐ C
☐ FA
☑ GA

Summary

Count:	86
AGG(COUNT([User ID])/COUNTD([User ID]))	
Sum:	233.653
Average:	2.717
Minimum:	1.593
Maximum:	8.915
Median:	2.334
Standard deviation:	1.246

- Here the average is 2.717 and median is 2.334
- Standard Deviation is 1.246

OUTLIERS:

- **In Type C, ACM_SIGHPC is an outlier with an edit average of 4.33 . This seems to be a non profit and the article isn't very detailed either.**

- **In FA, NeXT has a low average, this might be due to it being a defunct company associated with Steve Jobs.**
- **In GA, Development of Duke_Nukem_Forever is an outlier with low average, this is also an old game , which shouldn't have many edits in the first place**
- **In GA, Alan_Kotok stands out with a high average of 8.915. His contributions as a computer scientist, mean that there will be lot of citations and edits by people regarding his achievements**

CONCLUSION:

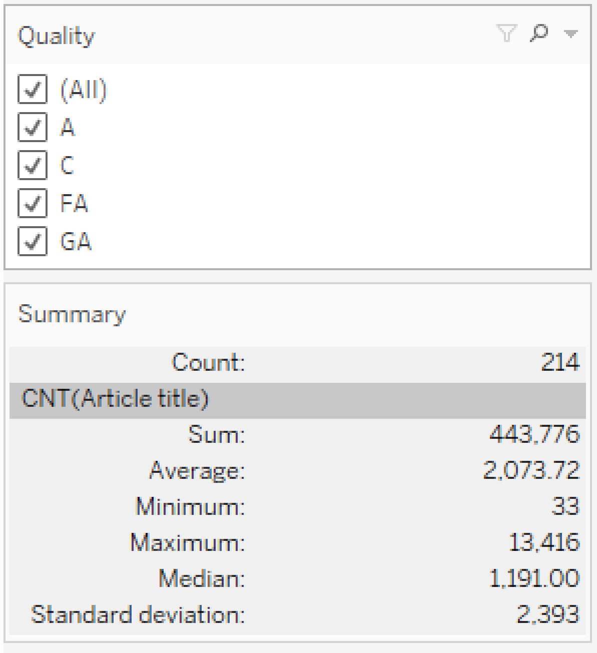
- **My hypothesis stands the test of data, except for the few outliers mentioned earlier**
- **Higher average edits per user seems to be better for the data sample included**

3. Total number of edits

HYPOTHESIS: Higher number of edits would mean lower quality of the article

- The column here is Article Quality
- The row here is the count (article Title) which will provide the number of instances that article name appeared in the data
- When article title is compared with this count, we can plot a box plot

Overall Summary:



Quality

☑ (All)
☑ A
☑ C
☑ FA
☑ GA

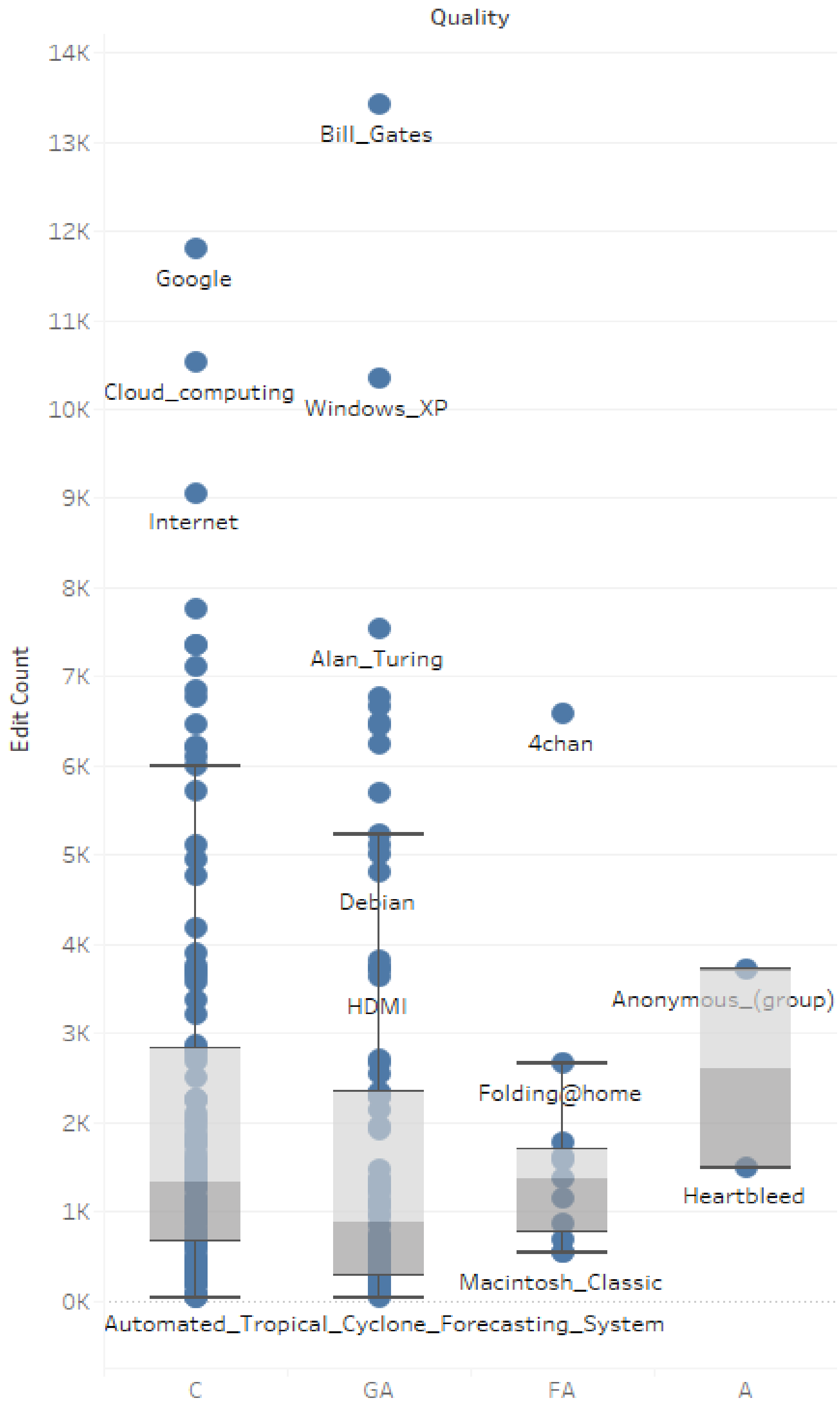
Summary

Count:	214
CNT(Article title)	
Sum:	443,776
Average:	2,073.72
Minimum:	33
Maximum:	13,416
Median:	1,191.00
Standard deviation:	2,393

- **Here average is 2073 overall and the median value is 1191**
- **The standard deviation is 2,393**

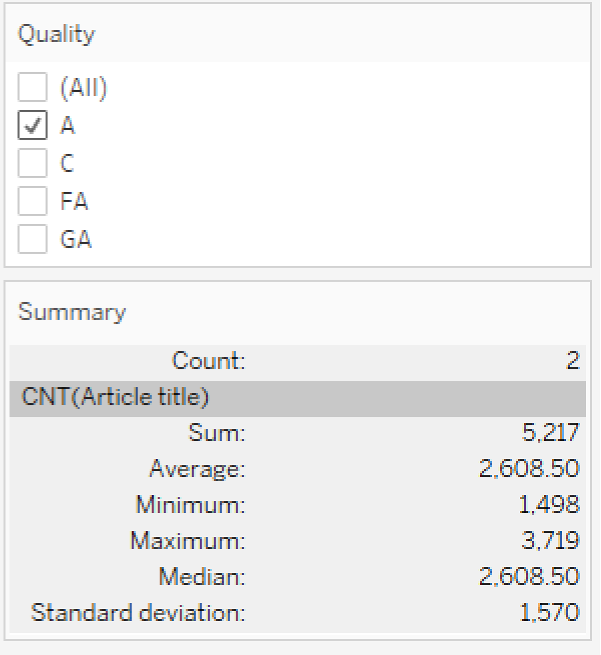
BOX PLOT

3.Total Number of Edits



Count of Article title for each Quality. The marks are labeled by Article title. Details are shown for Article title.

Type A



Quality

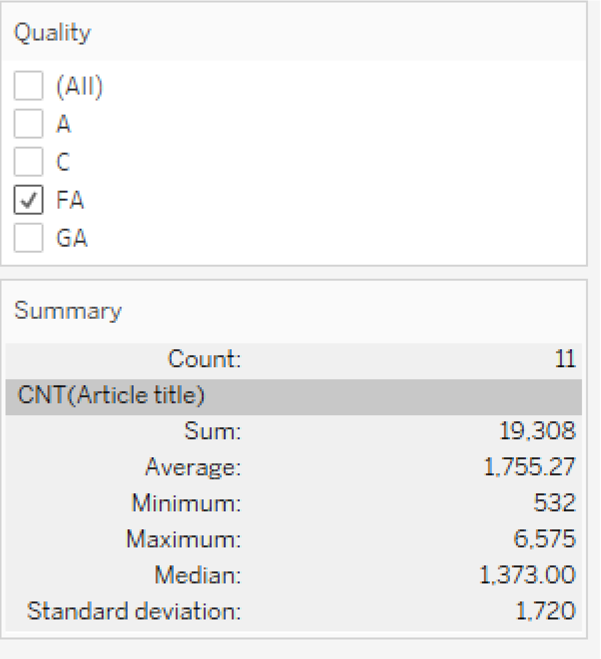
☐ (All)
☑ A
☐ C
☐ FA
☐ GA

Summary

Count:	2
CNT(Article title)	
Sum:	5,217
Average:	2,608.50
Minimum:	1,498
Maximum:	3,719
Median:	2,608.50
Standard deviation:	1,570

- Here there are only 2 types
- Average- 2,608.50 = Median
- Standard Deviation – 1570

TYPE FA



Quality

☐ (All)
☐ A
☐ C
☑ FA
☐ GA

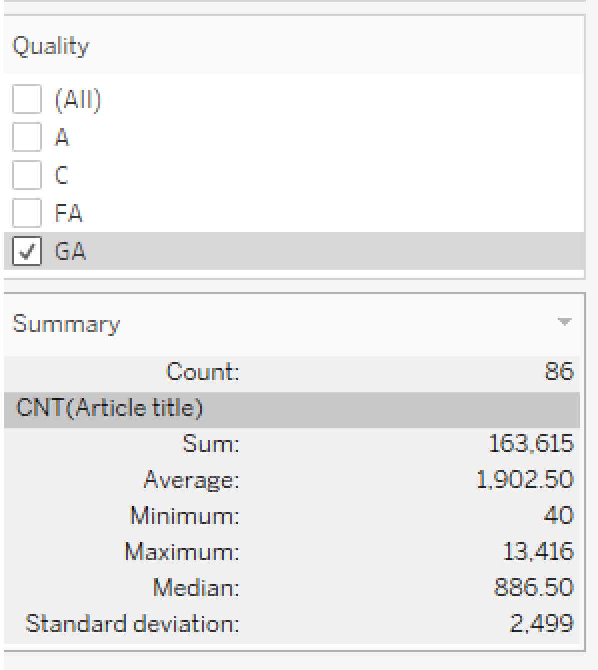
Summary

Count:	11
CNT(Article title)	
Sum:	19,308
Average:	1,755.27
Minimum:	532
Maximum:	6,575
Median:	1,373.00
Standard deviation:	1,720

Here the average is 1,755.27 while the median is 1373

Standard Deviation is 1720 for Type FA (Featured Articles)

TYPE GA



Quality

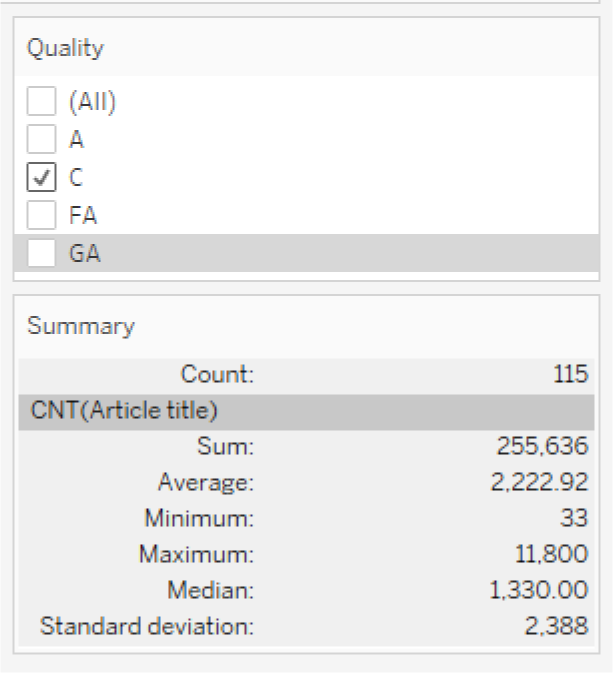
☐ (All)
☐ A
☐ C
☐ FA
☑ GA

Summary

```
              Count:        86
CNT(Article title)
                Sum:   163,615
            Average:  1,902.50
            Minimum:        40
            Maximum:    13,416
             Median:    886.50
 Standard deviation:     2,499
```

- The average is 1902.50 while the median is 886.50
- The standard deviation is higher here at 2,499

Type C(Bad)



Quality

☐ (All)
☐ A
☑ C
☐ FA
☐ GA

Summary

```
              Count:       115
CNT(Article title)
                Sum:   255,636
            Average:  2,222.92
            Minimum:        33
            Maximum:    11,800
             Median:  1,330.00
 Standard deviation:     2,388
```

- Here the average is 2292.92
- Here the median is 1330
- The standard deviation is 238

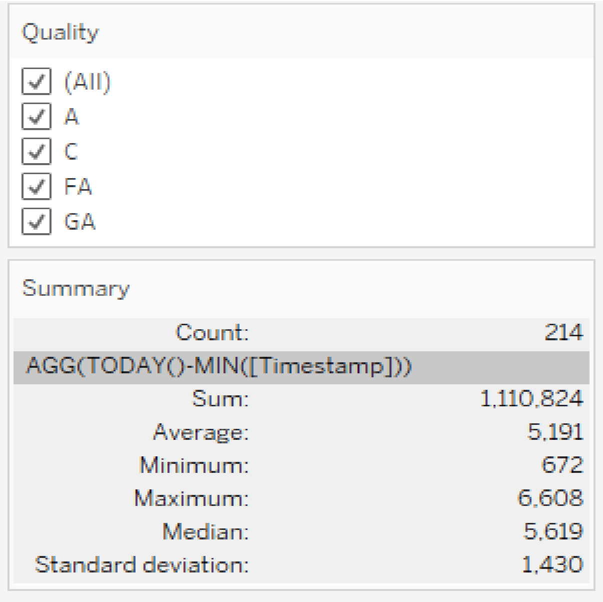
Outliers and Conclusions:

- **Overall, excluding Type A which has only 2 instances, higher average of user edits per article belongs to Type C which confirms my hypothesis**
- **Further, we can see some outliers as listed:**
    - **For types GA and FA, the articles such as Bill Gates,Windows_XP,Alan_Turing,4chan are all topics with lot of changes expected regularly hence higher number of edits**
    - **For type C, according to my hypothesis, topics such as ACM_sighPC,Cray_XMP have low edits as they are newer in the case of ACM_SighPC while Cray XMP is an old supercomputer based in 1980s**

4. Time since the first edit

- **Hypothesis:** Newer articles would have lower quality , So my hypothesis would be that Type C articles will be newer articles or the time difference will be lesser and hence will be of lower quality
- In Tableau, I did the following:
    - Column indicates Article Quality
    - In rows , I considered the Article Name as well as the **aggregate function which is the difference between the day of creation and today . This is measured in days.**
    - Formula : TODAY()-MIN([Timestamp])
    - Here I have exported into "4.TimeSinceFirstEdit.csv" with days mentioned in the column under quality

Overall Summary:



Quality

☑ (All)
☑ A
☑ C
☑ FA
☑ GA

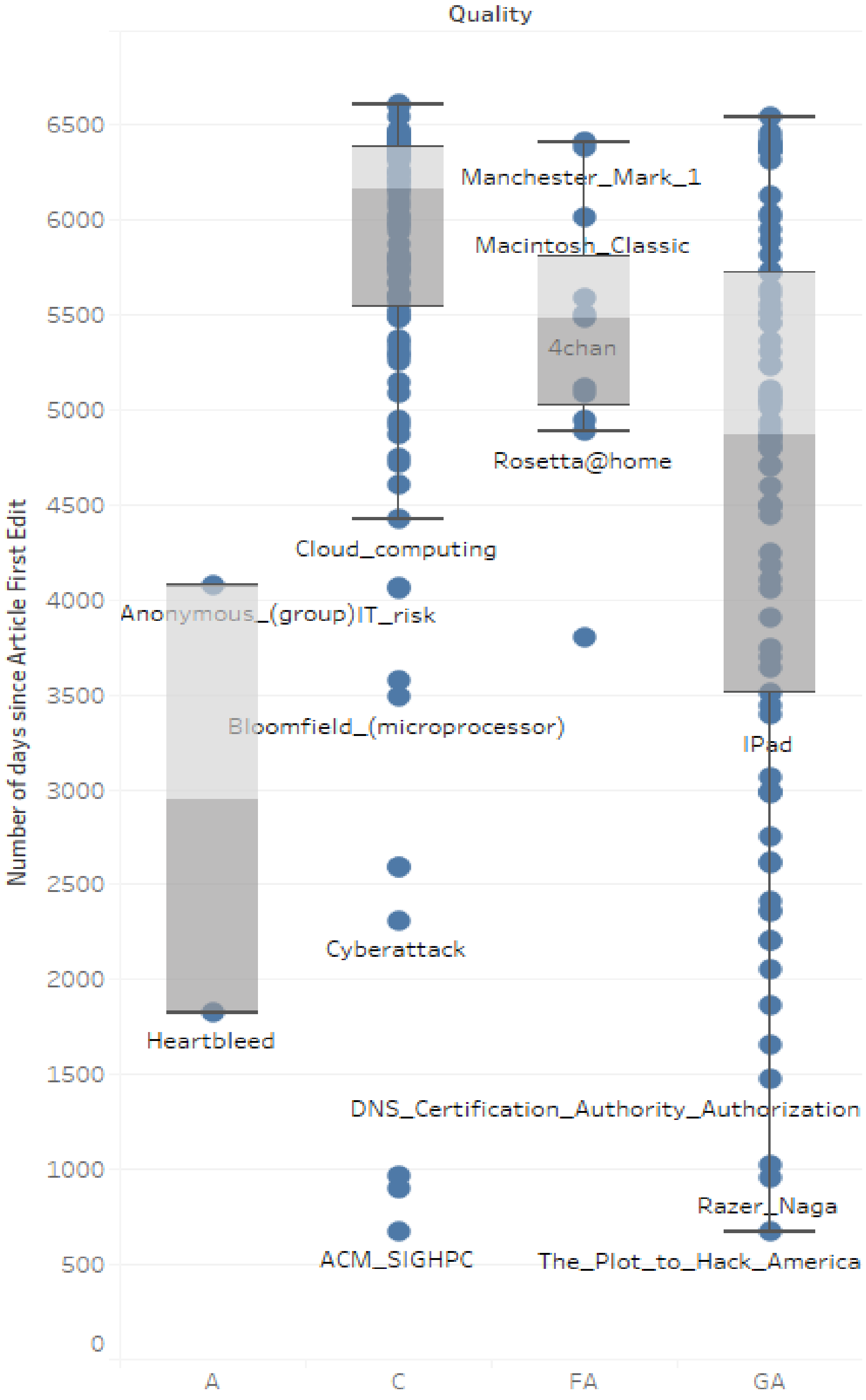
Summary

Count:	214
AGG(TODAY()-MIN([Timestamp]))	
Sum:	1,110,824
Average:	5,191
Minimum:	672
Maximum:	6,608
Median:	5,619
Standard deviation:	1,430

- The average number of days is 5191 for articles
- The median is 5619
- Standard deviation is around 1430

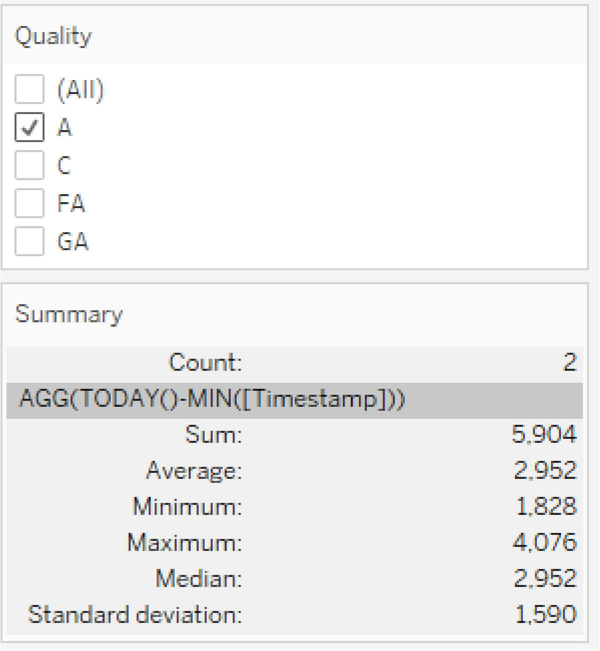
BOX PLOT

4.TimeSinceFirstEdit (2)



TODAY()-MIN([Timestamp]) for each Quality. The marks are labeled by Article title. Details are shown for Article title. The view is filtered on Quality, which keeps A, C, FA and GA.

TYPE A



Quality

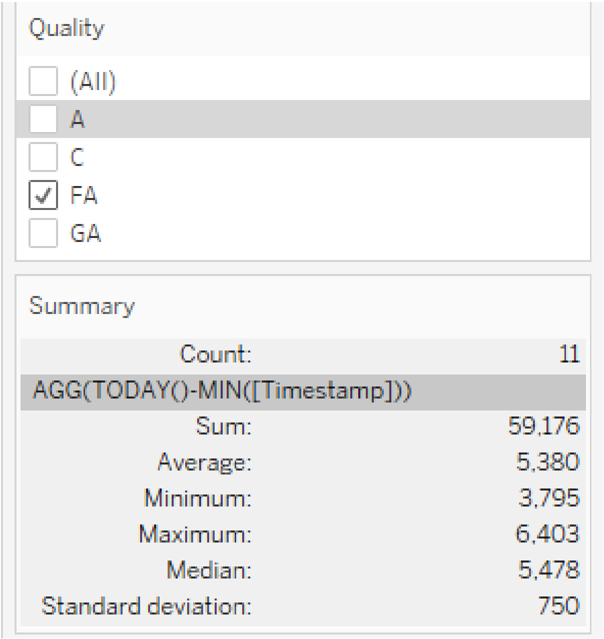
□ (All)
☑ A
□ C
□ FA
□ GA

Summary

Count:	2
AGG(TODAY()-MIN([Timestamp]))	
Sum:	5,904
Average:	2,952
Minimum:	1,828
Maximum:	4,076
Median:	2,952
Standard deviation:	1,590

- Here average is 2952 but there are only 2 cases in Type A which will be the same as the average
- Standard Deviation is 1590

TYPE FA



Quality

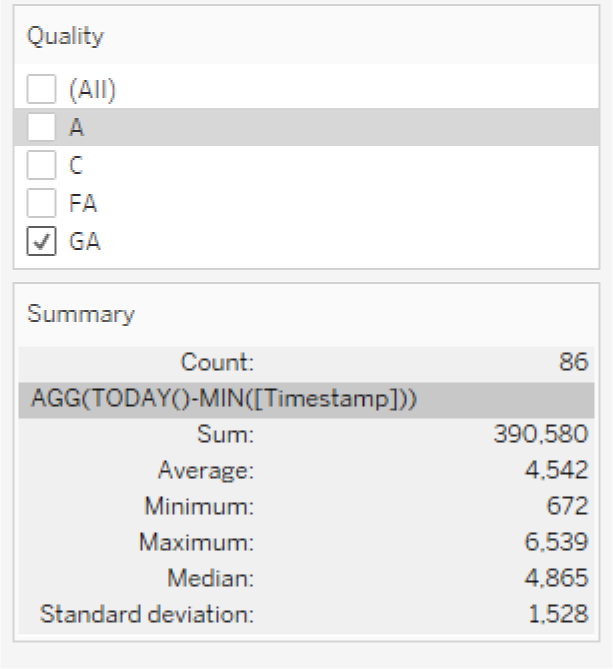
□ (All)
□ A
□ C
☑ FA
□ GA

Summary

Count:	11
AGG(TODAY()-MIN([Timestamp]))	
Sum:	59,176
Average:	5,380
Minimum:	3,795
Maximum:	6,403
Median:	5,478
Standard deviation:	750

- Here there are 11 cases
- The average is 5380
- The Median is 5478
- Standard deviation :750

Type GA:



Quality

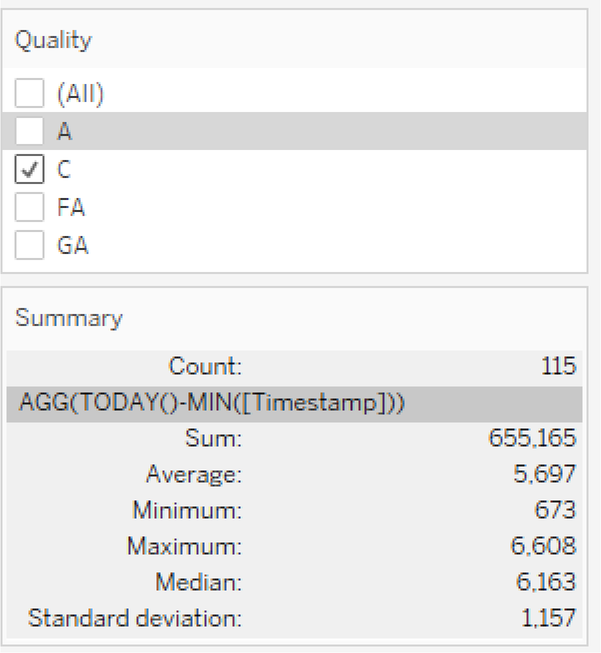
☐ (All)
☐ A
☐ C
☐ FA
☑ GA

Summary

Count:	86
AGG(TODAY()-MIN([Timestamp]))	
Sum:	390,580
Average:	4,542
Minimum:	672
Maximum:	6,539
Median:	4,865
Standard deviation:	1,528

- The average number of days is 4542 while median stands at 4865
- The Standard deviation is 1528 for GA

TYPE C(Bad)



Quality

☐ (All)
☐ A
☑ C
☐ FA
☐ GA

Summary

Count:	115
AGG(TODAY()-MIN([Timestamp]))	
Sum:	655,165
Average:	5,697
Minimum:	673
Maximum:	6,608
Median:	6,163
Standard deviation:	1,157

- Average case here is 5697 while median is 6163
- The standard deviation is 1157

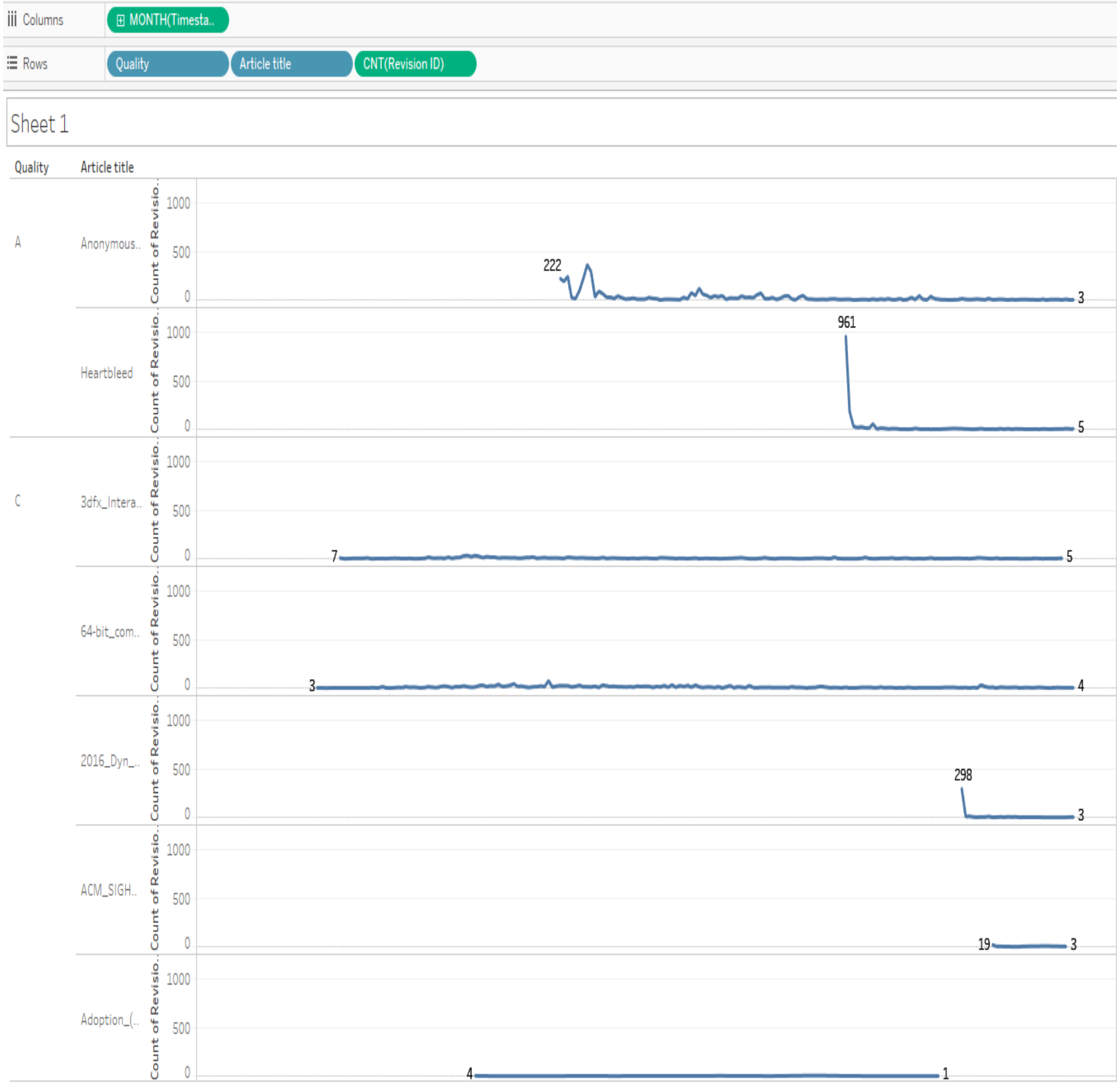
Outliers and Conclusions:

- **My HYPOTHESIS based on the sample seems to be *WRONG* as the average is higher for type C . This could be due to a good chunk of articles that are newer in type GA due to varying reasons.**
- **Outliers:**
    - **In Type C: Central Processing Unit, Bit,IEEE 802.11,Alan_Key are really old and related to computer science in some way or the other. On the other end of the spectrum, ACM_SighPC and Dell Technologies are newer. Dell technologies is the new name of Dell**
    - **In Type GA: The_plot to hack America ,Razer_Naga and Qapital are newer but good articles. The first 2 read more like long reviews so it makes sense that they are rated well. Qapital has decent number of references but prima facie , nothing really stands out there**

5. **Number of Edits in First Month:**

Hypothesis: More the number of edits in the first month, the better the article quality

- This was the most difficult one
- In Tableau , I created a dataset where:
    - Column for Month(Timestamp)
    - Row with Quality,Article Title,Count(Revision ID)
    - With this I got a graph where I could see the number of revisions per month for every article in each row (Snapshot in next page)
    - I couldn't get only first month, tried rank and other formulas for a few hours
    - Next, I made an excel file after exporting only 2 columns, article name and quality
    - In this step , I manually entered the data (This was time consuming)
    - I have attached multiple files, showing what all I tried including the final file called "55.xlsx"
-



Columns: MONTH(Timesta..

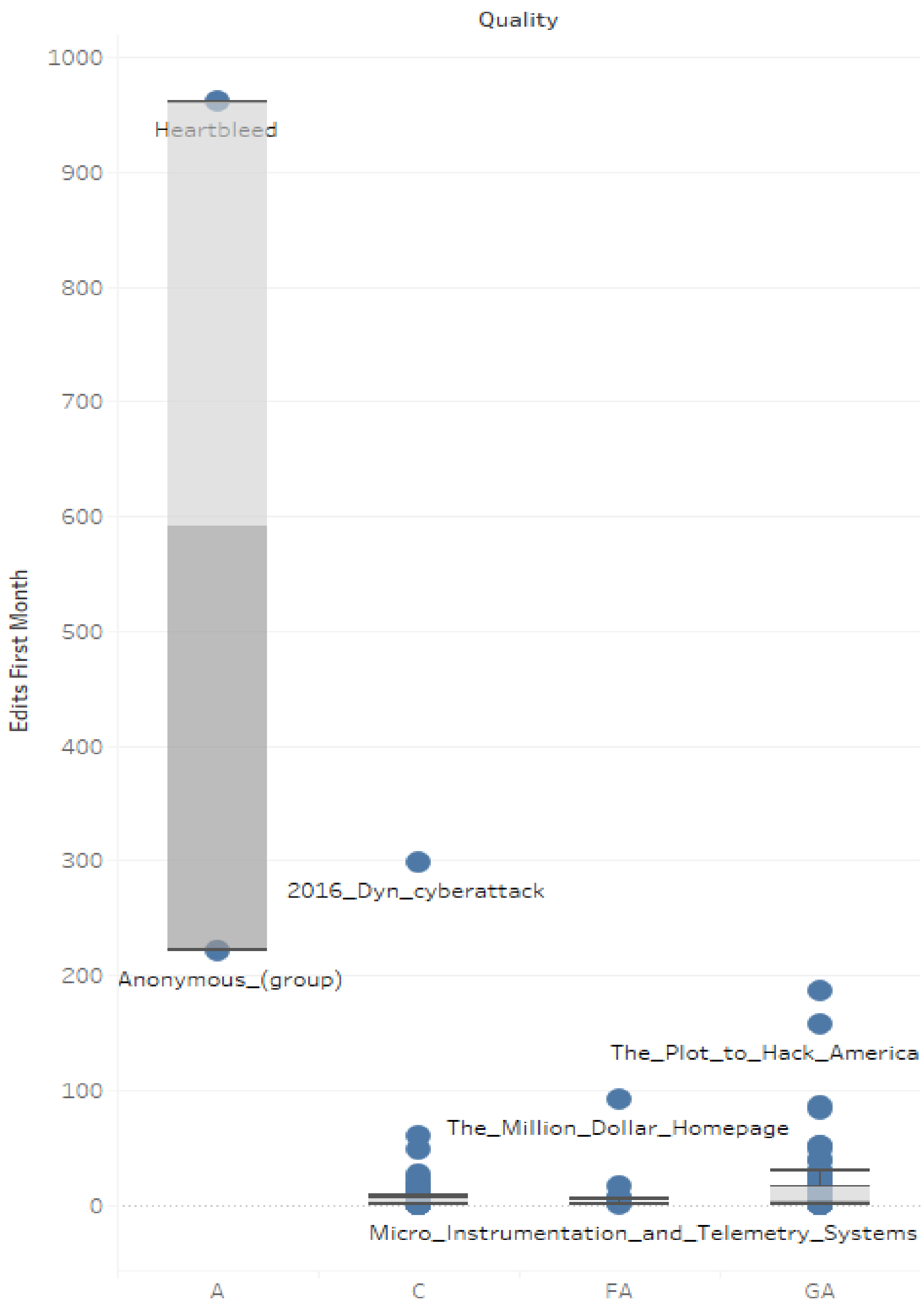
Rows: Quality | Article title | CNT(Revision ID)

Sheet 1

Figure: Snapshot of Tableau Work for Question 5

BOX PLOT

Sheet 1



Edits First Month for each Quality. The marks are labeled by Article title. Details are shown for Article title. The view is filtered on Quality, which keeps A, C, FA and GA.

For Type A

- Median : 591.5
- Average :591.5
- Standard Deviation:369.5

TYPE C

- Average – 7.19
- Median :2
- Standard Deviation :3

Type FA:

- Average: 12.909
- Median : 4
- Standard Deviation 2.5

TYPE GA

- Average : 14.54
- Median 4

Conclusions and Outliers:

- **Let's ignore Type A as it has only 2 values.**
- **But for the others, we see that Type C has a lower average as per the hypothesis compared to type FA and Type GA**
- **Therefore initial suspicions are confirmed**
- **In regards to outliers:**
    - **2016_Dyn_CyberAttack has a high average – This might be due to when the event struck , the details where entered in.**
    - **In FA and GA cases:**
        - **Manchester_Mark_1 : this is an old 1949 computer, so it was already probably a detailed article**
        - **Micro_Instrumentation_and_Telemetry_Systems: This was also detailed in the first place.**

CONCLUSION OVERALL

- After considering all 5 scenarios, the following can be said:
  - Higher number of overall edits on an article tends to indicate poorer quality or ‘C’ level (Generally) with some exceptions
  - The higher the unique editors, the chances of the article being poor increase
  - Not much can be said about ‘A’ level articles as there are only 2 of them
  - The higher the average edits per user, the better the article. This seems to be correct as well
  - The older the article, the chances of it being poor increases or atleast that is the case with the sample data
  - Higher edits in the first month means a better chance of the article being good
  - The best indicator overall:
    - Higher number of average edits

-