Data Insights Report

Objectives

The objectives of this case study are:

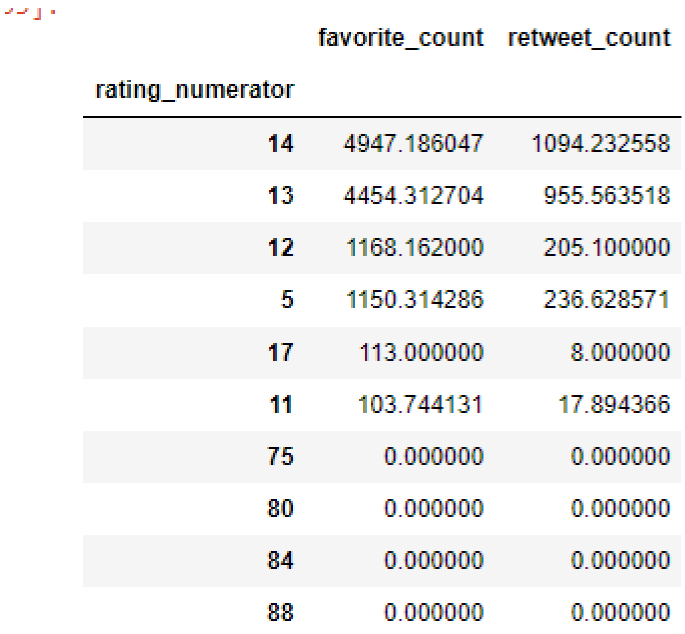
- **Analyzing**: The clean data were explored using programmatic assessments, to gain insights and trends from the dataset.
- **Visualizing**: After analyzing the data, charts were created to engage the audience on the insights and trends derived from the analysis.
- **Conclusion**: Finally this case study was ended with conclusions based on the insights derived from the data

Analyzing

In this phase programmatic assessments were used to gain insights and trends from the dataset. We were able to make bivariate comparison to find correlation and derive desired output from summary statistics.

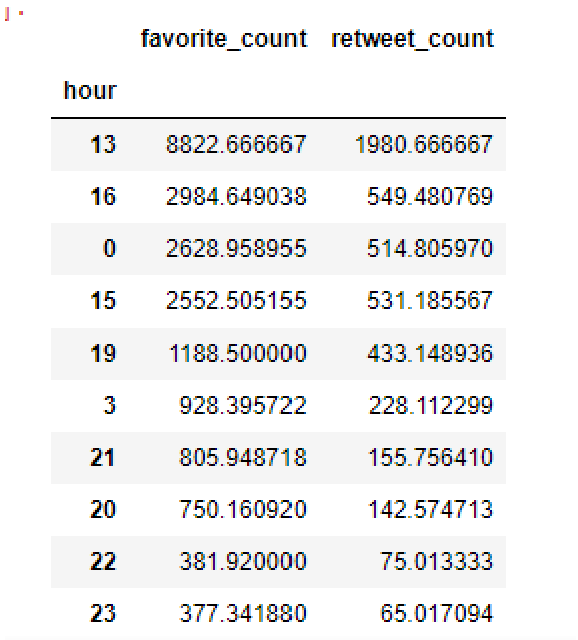
These were the insights gained from the analysis of 'twitter_archived_clean' table:

- Rating number '14' had the highest average favorite count and retweet count followed by rating number `'13'. From visual assessment we see that there is a positive correlation between favorite count and retweet count



	favorite_count	retweet_count
rating_numerator		
14	4947.186047	1094.232558
13	4454.312704	955.563518
12	1168.162000	205.100000
5	1150.314286	236.628571
17	113.000000	8.000000
11	103.744131	17.894366
75	0.000000	0.000000
80	0.000000	0.000000
84	0.000000	0.000000
88	0.000000	0.000000

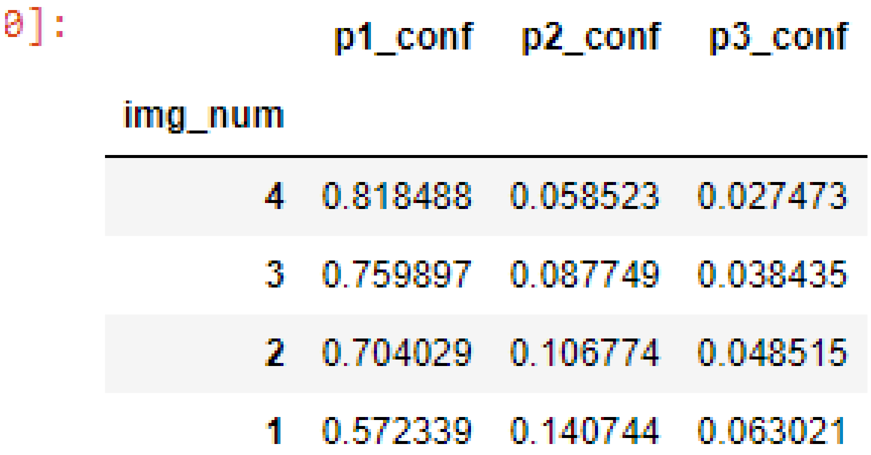
- we can denote that tweets at the range of 13:00PM tends to have a high average favorite count and retweet count, and from visual assessment we can denote that engagements(favorite count and retweet count) are at a high end from 13.00PM and from 4:00AM to 11:00AM it's at a low end



]:

	favorite_count	retweet_count
hour		
13	8822.666667	1980.666667
16	2984.649038	549.480769
0	2628.958955	514.805970
15	2552.505155	531.185567
19	1188.500000	433.148936
3	928.395722	228.112299
21	805.948718	155.756410
20	750.160920	142.574713
22	381.920000	75.013333
23	377.341880	65.017094

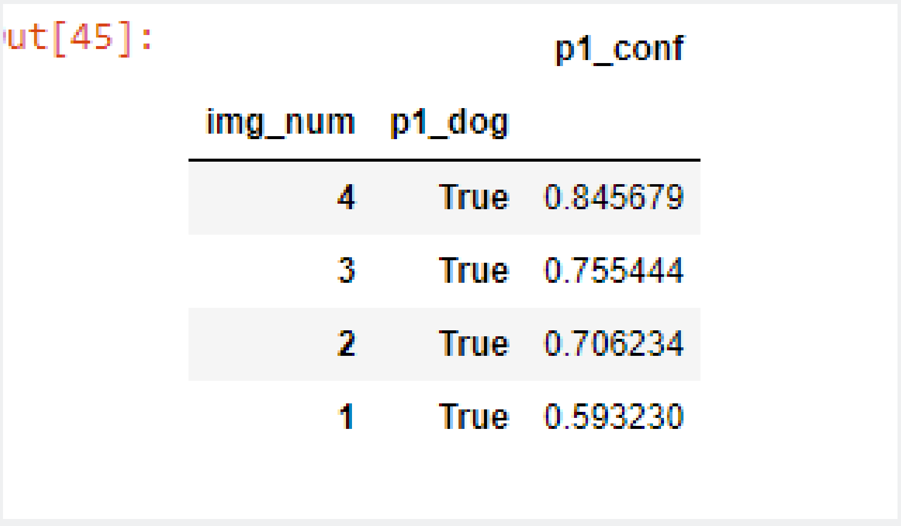
- Image 4 has the highest average confident prediction for p1_conf,image 1 has the highest average confident prediction for p2_conf and p3_conf respectively.



0]:

	p1_conf	p2_conf	p3_conf
img_num			
4	0.818488	0.058523	0.027473
3	0.759897	0.087749	0.038435
2	0.704029	0.106774	0.048515
1	0.572339	0.140744	0.063021

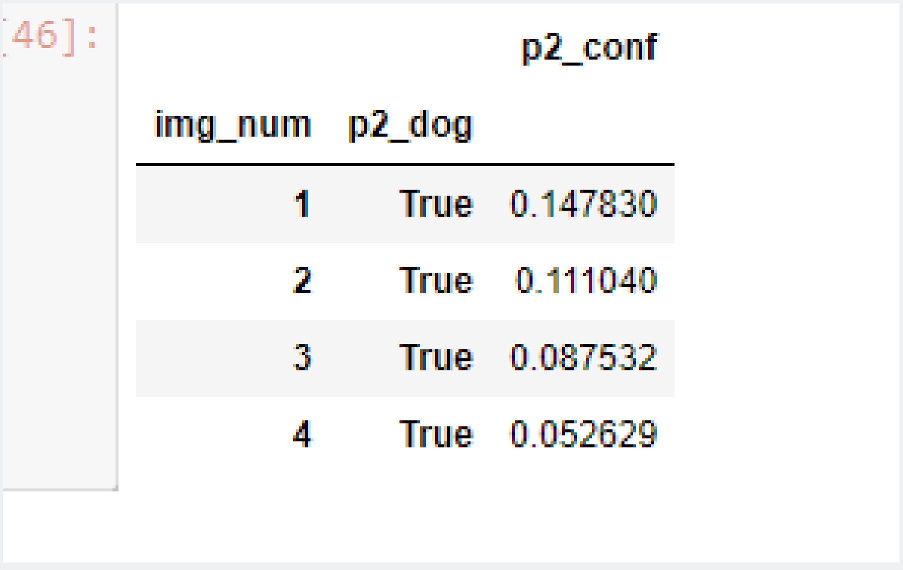
- We can denote that after filtering out non-breeds of dogs, image number 4 tends to have the highest average confidence prediction for p1_conf



ut[45]:

		p1_conf
img_num	p1_dog	
4	True	0.845679
3	True	0.755444
2	True	0.706234
1	True	0.593230

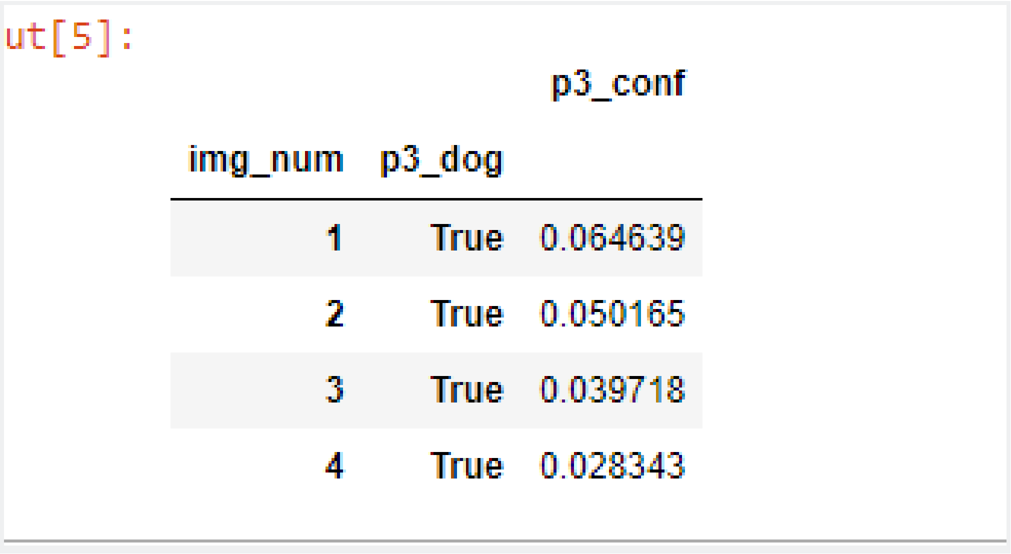
- So we can denote that after filtering out non-breeds of dogs, image number 1 tends to have the highest average confidence prediction for p2_conf



[46]:

		p2_conf
img_num	p2_dog	
1	True	0.147830
2	True	0.111040
3	True	0.087532
4	True	0.052629

- So we can denote that after filtering out non-breeds of dogs, image number 1 tends to have the highest average confidence prediction for p3_conf



ut[5]:

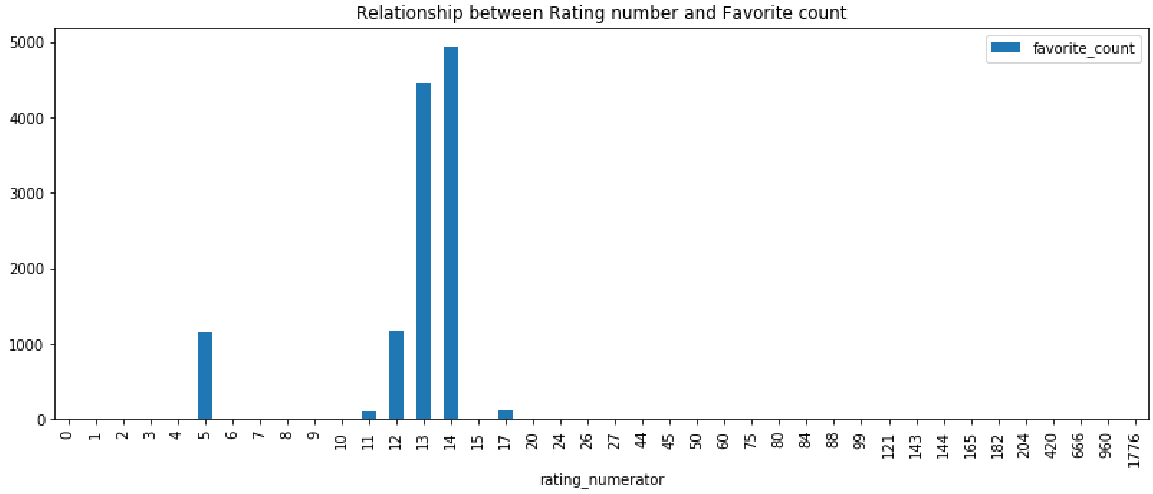
		p3_conf
img_num	p3_dog	
1	True	0.064639
2	True	0.050165
3	True	0.039718
4	True	0.028343

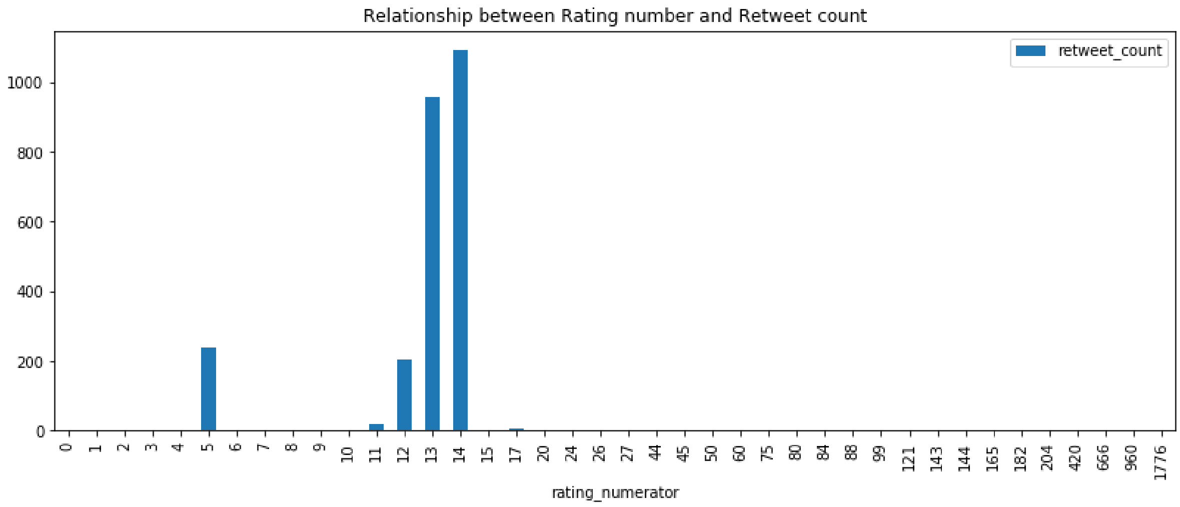
Visualization

In this phase data is visualized using charts and graphs to explain the insights gained from the data.

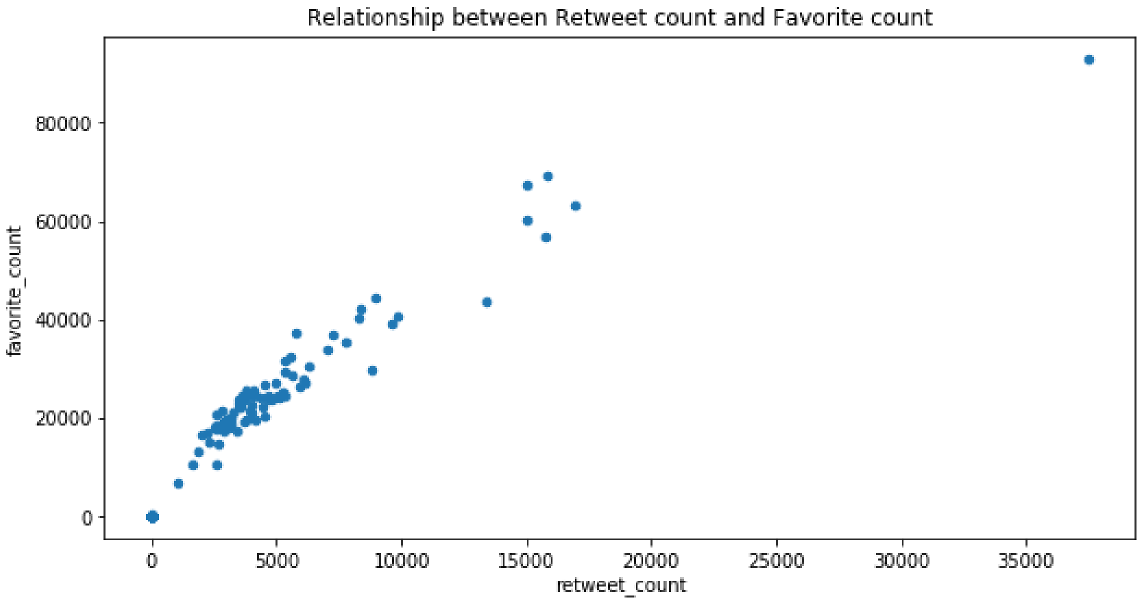
These were the charts derived from manipulating the wrangled data of 'twitter_archived' table:

- Rating 14 had the highest average favorite_count and retweet_count followed by rating 13.

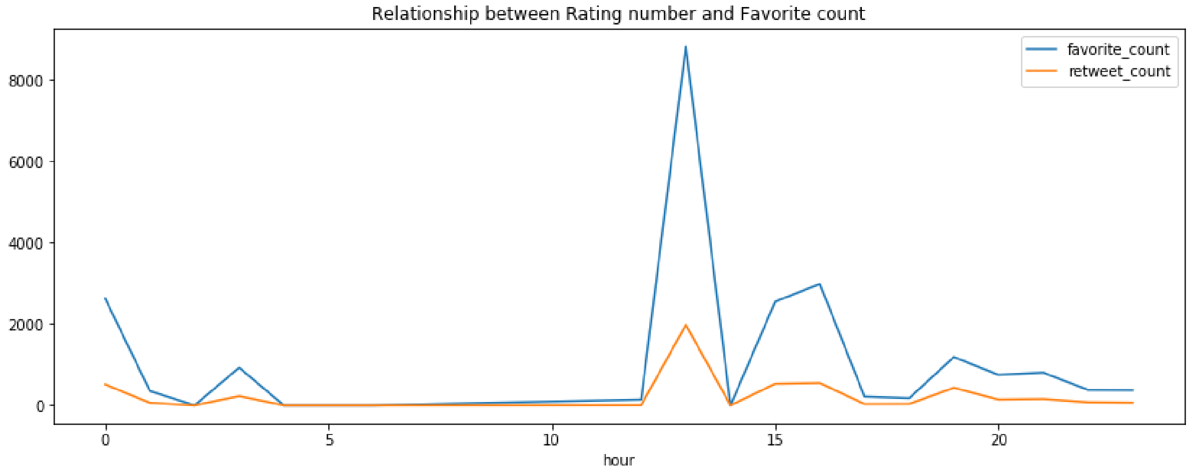




- From visual assessment we see that there is a positive correlation between favorite count and retweet count

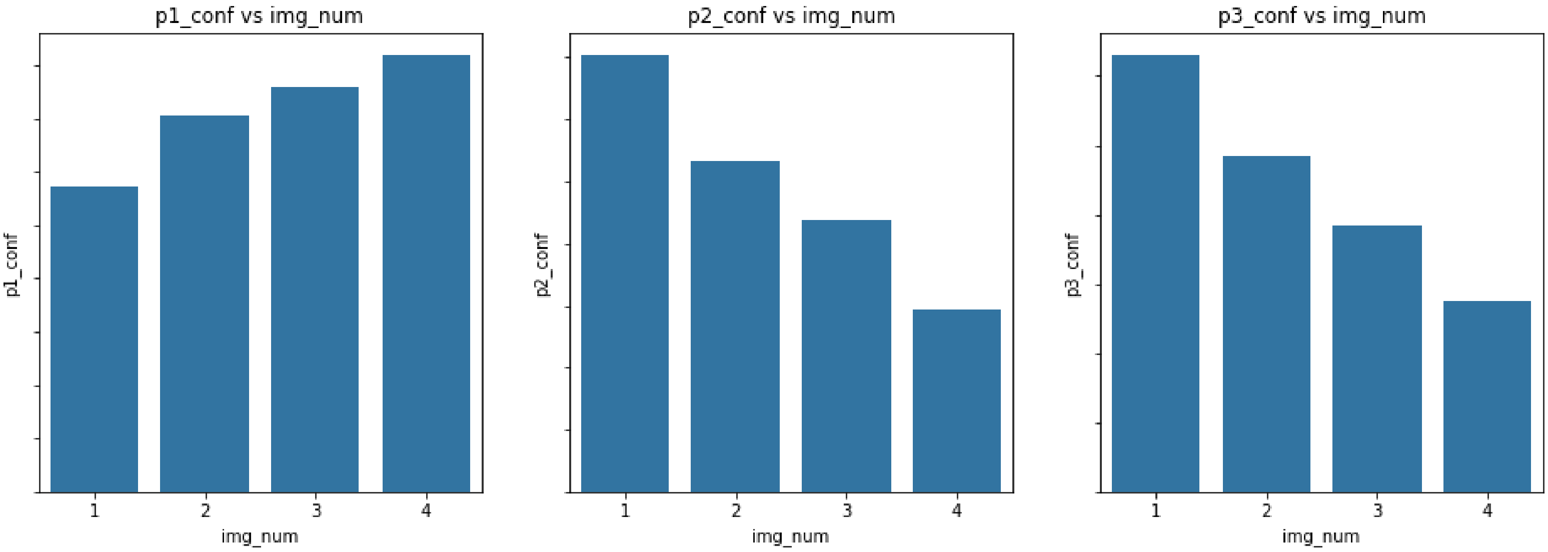


- From visual assessment we can denote that engagements(favorite count and retweet count) are at a peak at 13.00PM and uninteractive hours from 4:00AM to 11:00AM

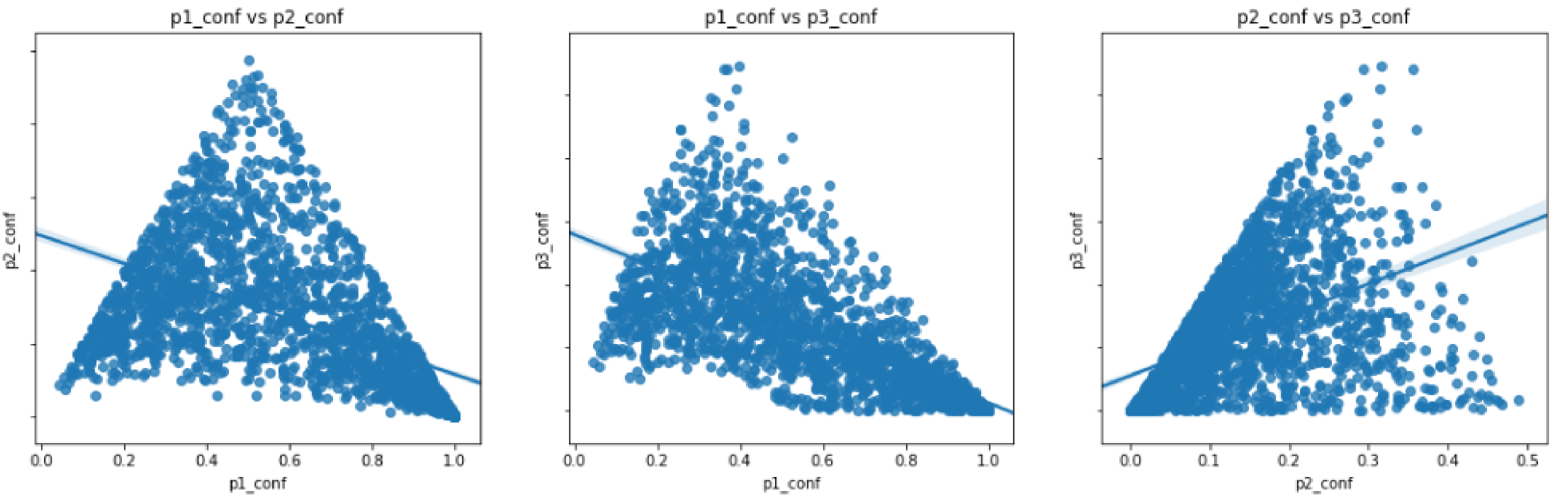


These were the charts derived from manipulating the wrangled data of 'twitter_archived_clean' table:

- Image 4 has the highest average confident prediction for p1_conf,image 1 has the highest average confident prediction for p2_conf and p3_conf respectively.



- There is a negative correlation in first chart(p1_conf and p2_conf), negative correlation in second chart(p1_conf and p3_conf), postive correlation in third chart(p3_conf and p2_conf)



Conclusions

- Tweets with a rating of 13 and 11 drew more engagements(favorite_count and retweet_count) than the rest and favorite count has a positive correlation with retweet which means the more the favorite count the more the retweet count.
- Hour of the day had an impact on the number of engagements(favorite_count and retweet_count), with 13:00PM being the peak of engagements and hours earlier tends to draw fewer engagements and than the later hours from 13:00AM
- Image number and confident prediction has a correlation which means that for the first picture out of the top three, image number 4 has a average rating confident prediction of 82% with the remaining 18% shared amongst the remaining picture numbers in a descending order. Image number 3 has a average rating confident prediction of 76% with the remaining 24% shared amongst the remaining picture numbers in a descending order. Image number 2 has a average rating confident prediction of 70% with the remaining 30% shared amongst the remaining picture numbers in a descending order and image number 1 has a average rating confident prediction of 57% with the remaining 43% shared amongst the remaining picture numbers in a

descending order. Image number 1 tends to have the highest average confidence prediction whenever it’s not the first picture number.

- Since p1_conf > p2_conf>p3_conf there is a negative correlation in first chart(p1_conf and p2_conf), negative correlation in second chart(p1_conf and p3_conf), postive correlation in third chart(p3_conf and p2_conf)