Incident Prediction – Initial analysis

BY –

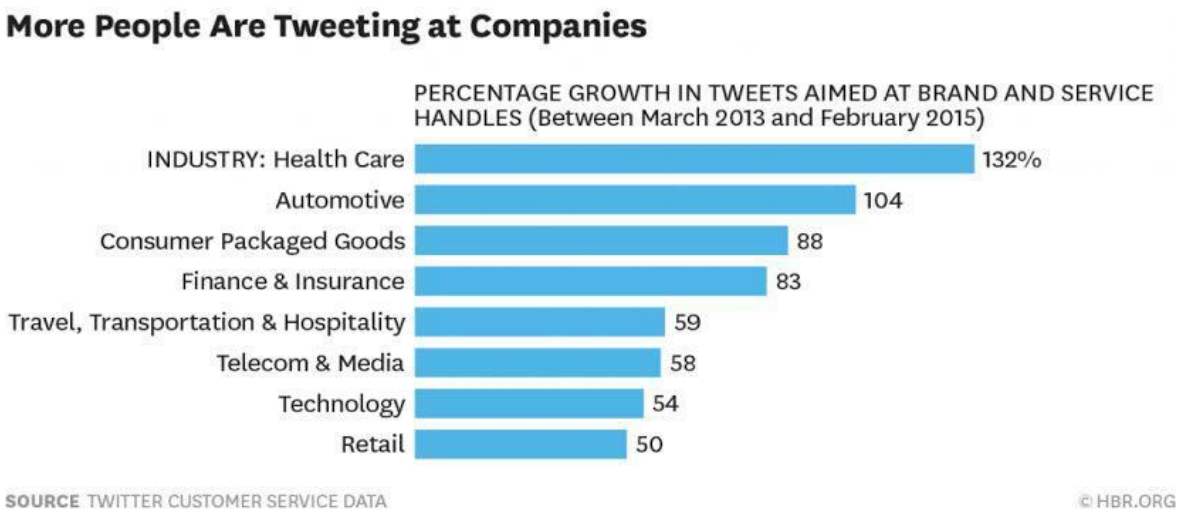
VAISHNAVI EM, RASHMI KULKARNI, SHRADDHA MANKAR, SUCHARITA MUKHERJEE

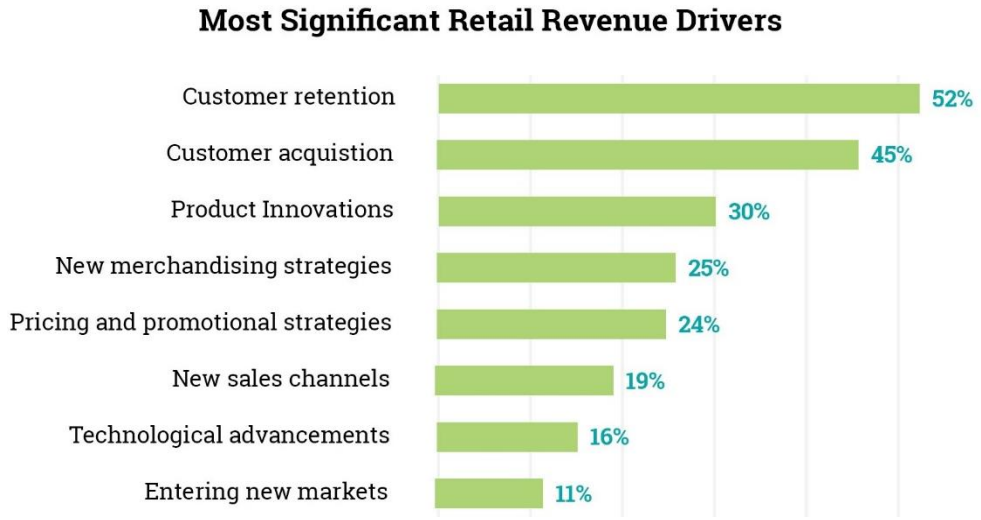
Background Study

Increasing digital footprint of almost every product and service industry is leading to a huge digital data load. The same data can be used for a better CRM(Customer relationship management)

With advent and easy access of multiple social media and other platforms its imperative that people are sharing feedbacks-grievances and more at a much higher pace.

Along with genuine concerns industries today also face a lot of scam-fake-negative-repetitive feedbacks-grievances, so it is importance to device a solution that can prioritise important concerns and can facilitate a better **customer satisfaction** leading to better **customer acquisition.**





Business Objective

To predict the impact of the incident raised by the customer by **prioritising the incident** as high-medium-low priority.

The dataset is having incidents raised by customers. Which contains an event log of an incident management process extracted from a service desk platform of an IT company.

Incident management workflow

(*reference: https://searchitoperations.techtarget.com/definition/IT-incident-management)*

Data Acquisition

The event log is enriched with **data loaded from a relational database** underlying a corresponding process-aware information system.

Dimension of the data is **1,41,712 incidents(rows) and 25 attributes(columns)**.

The attribute details are as follows:

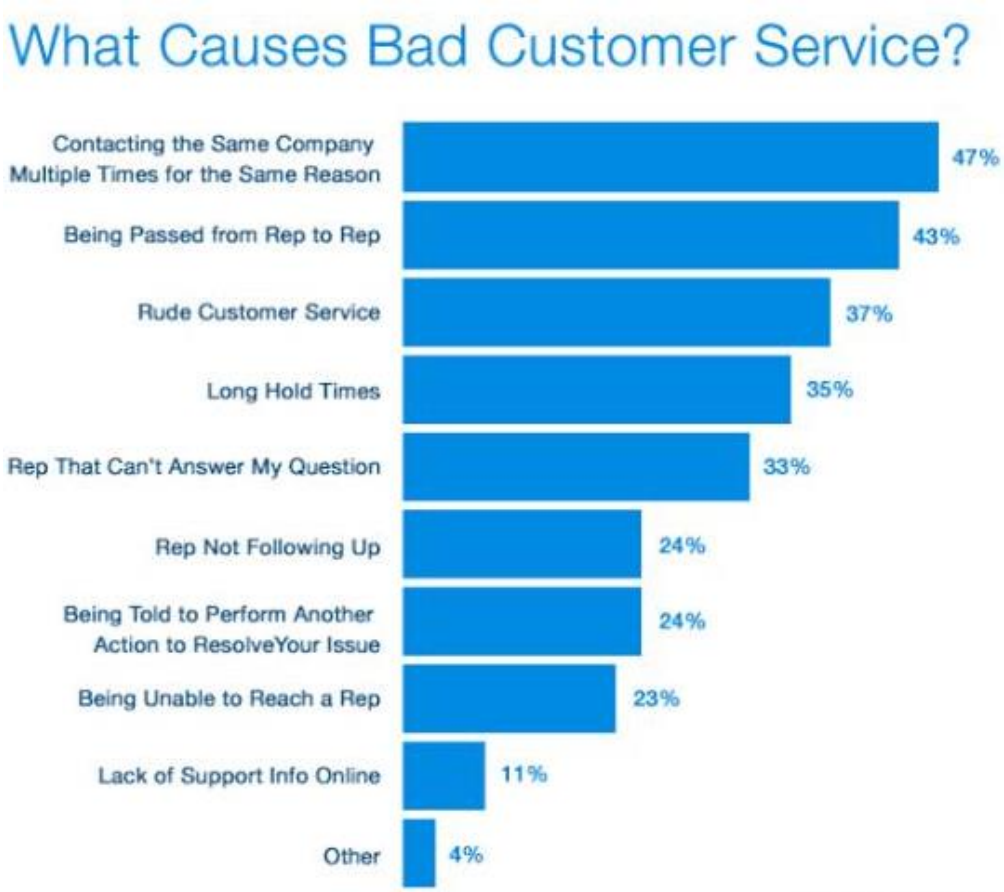
1. ID : Incident identifier (**24,918 different values**)
2. ID_status : Eight levels controlling the incident management process transitions from opening until closing the case
3. Active : Boolean attribute that shows whether the record is active or closed/canceled
4. count_reassign : Number of times the incident has the group or the support analysts changed
5. count_opening : number of times the incident resolution was rejected by the caller
6. count_updated : number of incident updates until that moment
7. ID_caller : identifier of the user affected

Data Acquisition contd.. *(getting hold of useful information)*

1. opened_by : identifier of the user who reported the incident
2. opened_time : Incident user opening date and time
3. Created_by : identifier of the user who registered the incident
4. created_at : incident system creation date and time
5. updated_by : identifier of the user who updated the incident and generated the current log record
6. updated_at : incident system update date and time
7. type_contact : categorical attribute that shows by what means the incident was reported
8. location : identifier of the location of the place affected
9. Category Id : first-level description of the affected service
10. user_symptom : description of the user perception about service availability

Data Acquisition contd.. *(getting hold of useful information)*

1. Impact - description of the impact caused by the incident (values: “1:High” ; “2:Medium” ; “3:Low”)
2. Support_group - identifier of the support group in charge of the incident
3. support_incharge - identifier of the user in charge of the incident
4. Doc_knowledge - boolean attribute that shows whether a knowledge base document was used to resolve the incident
5. confirmation_check - boolean attribute that shows whether the priority field has been double-checked
6. Notify - Categorical attribute that shows whether notifications were generated for the incident
7. Problem_id - identifier of the problem associated with the incident
8. change_request - identifier of the change request associated with the incident;



Understanding the collected information

Data types – Object (16), Integer(3), Date & Time (3), Bool (3),

Object	Object	Integer	Date & Time	Boolean
ID	user_symptom	count_reassign	opened_time	Active
ID_Status	**IMPACT (target)**	count_opening	created_at	Doc_knowledge
ID_caller	Support_group	count_updated	updated_at	confirmation_check
opened_by	support_incharge			
Created_by	notify			
updated_by	problem_id			
type_contact	Change request			
location				
category_ID				

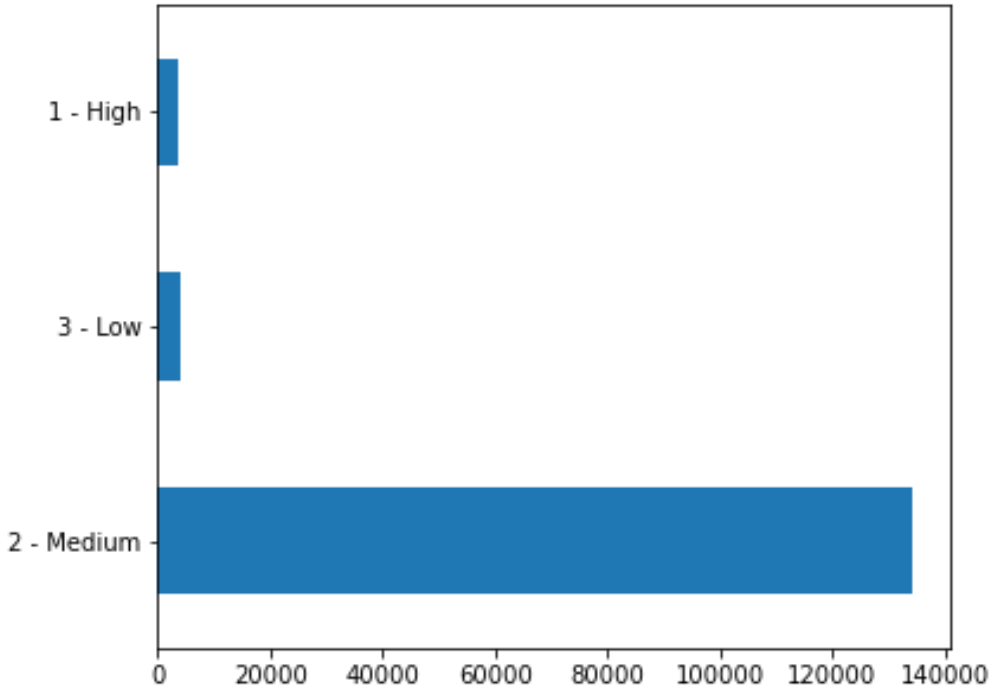
Understanding the distribution of the data

Study of skewness and kurtosis of the numerical and Boolean data values

Attribute	Skewness	Kurtosis	Interpretation
Active	-1.7	0.9	Negatively skewed data, there is long tail in the beginning, which comprise majority of the data
Count_reassign	3.1	16.5	Positively skewed data, there is long tail at the end, which comprise majority of the data
Count_opening	15.6	344.3	Positively skewed data, there is long tail at the end, which comprise majority of the data
Count_updated	4.7	35.4	Positively skewed data, there is long tail at the end, which comprise majority of the data
Doc_knowledge	1.7	0.8	Positively skewed data, there is long tail at the end, which comprise majority of the data
Confirmation_check	0.9	-1.1	Positively skewed data, there is long tail at the end, which comprise majority of the data

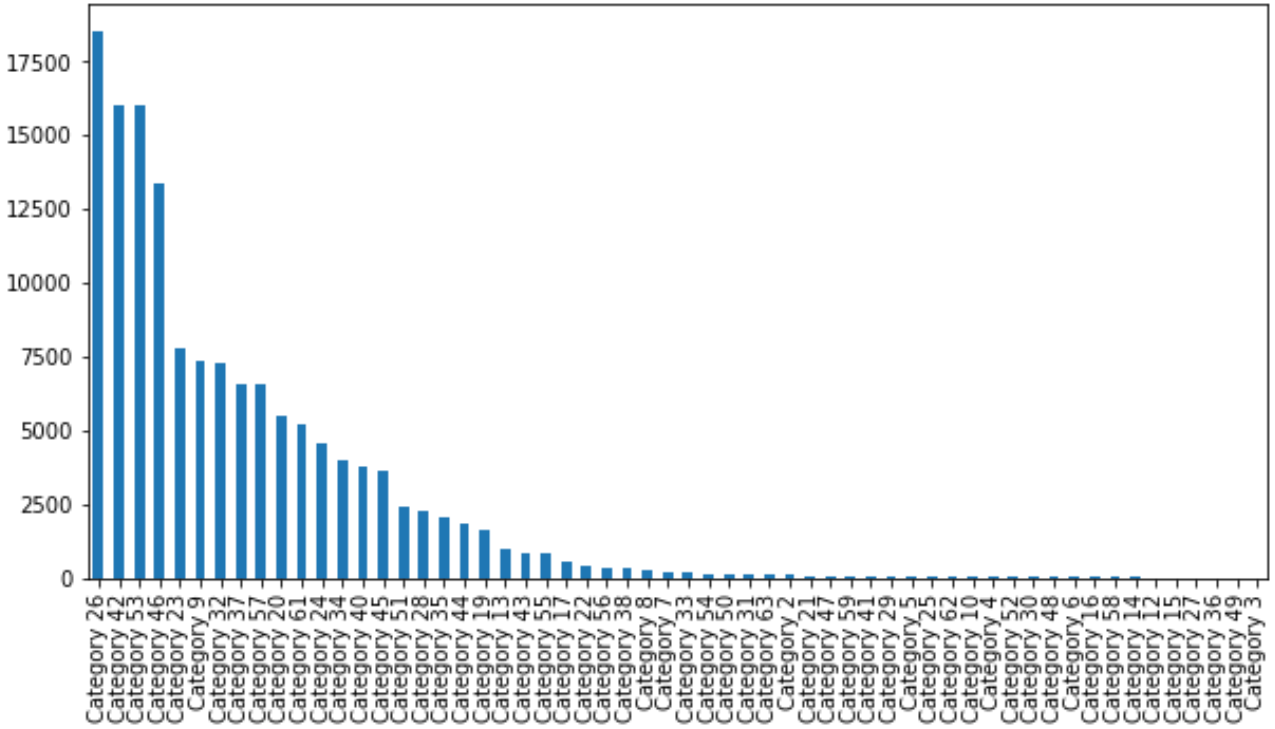
The data distribution is not symmetrical and does not follow the normal distribution patter. There are certain high and multiple low frequency points in the dataset

Visualizing the target distribution



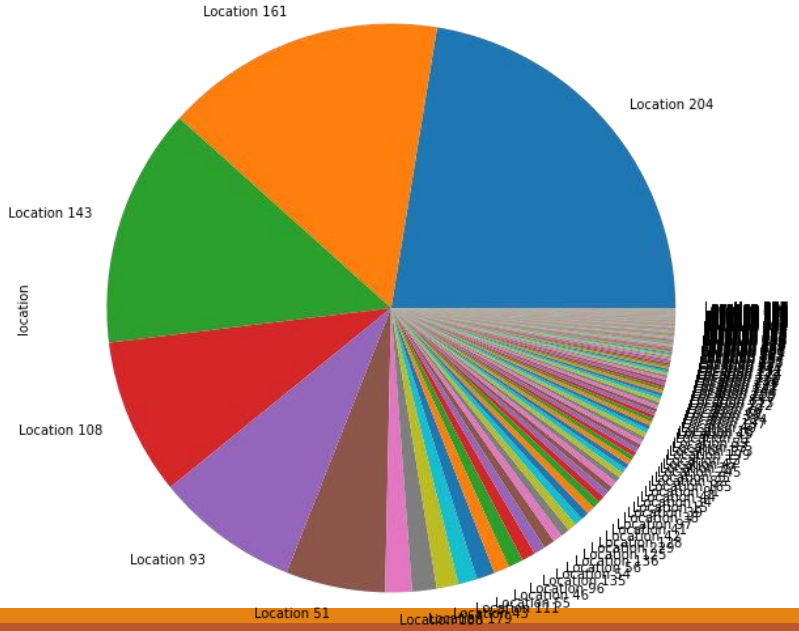
The data set is highly imbalanced with: 3491 High priority incidents, 134335 Medium priority incidents and 3886 Low priority incidents. Indicating a **highly imbalanced** dataset.

Visualizing few of the predictors



"ID_status" and "Impact"

ID_status	impact 1 - High	2 - Medium	3 - Low
-100		100.00%	
Active	2.48%	94.91%	2.61%
Awaiting Evidence	10.53%	84.21%	5.26%
Awaiting Problem	6.94%	84.60%	8.46%
Awaiting User Info	1.60%	95.15%	3.24%
Awaiting Vendor	3.39%	96.04%	0.57%
Closed	1.69%	95.29%	3.02%
New	3.39%	94.27%	2.34%
Resolved	2.25%	94.85%	2.90%

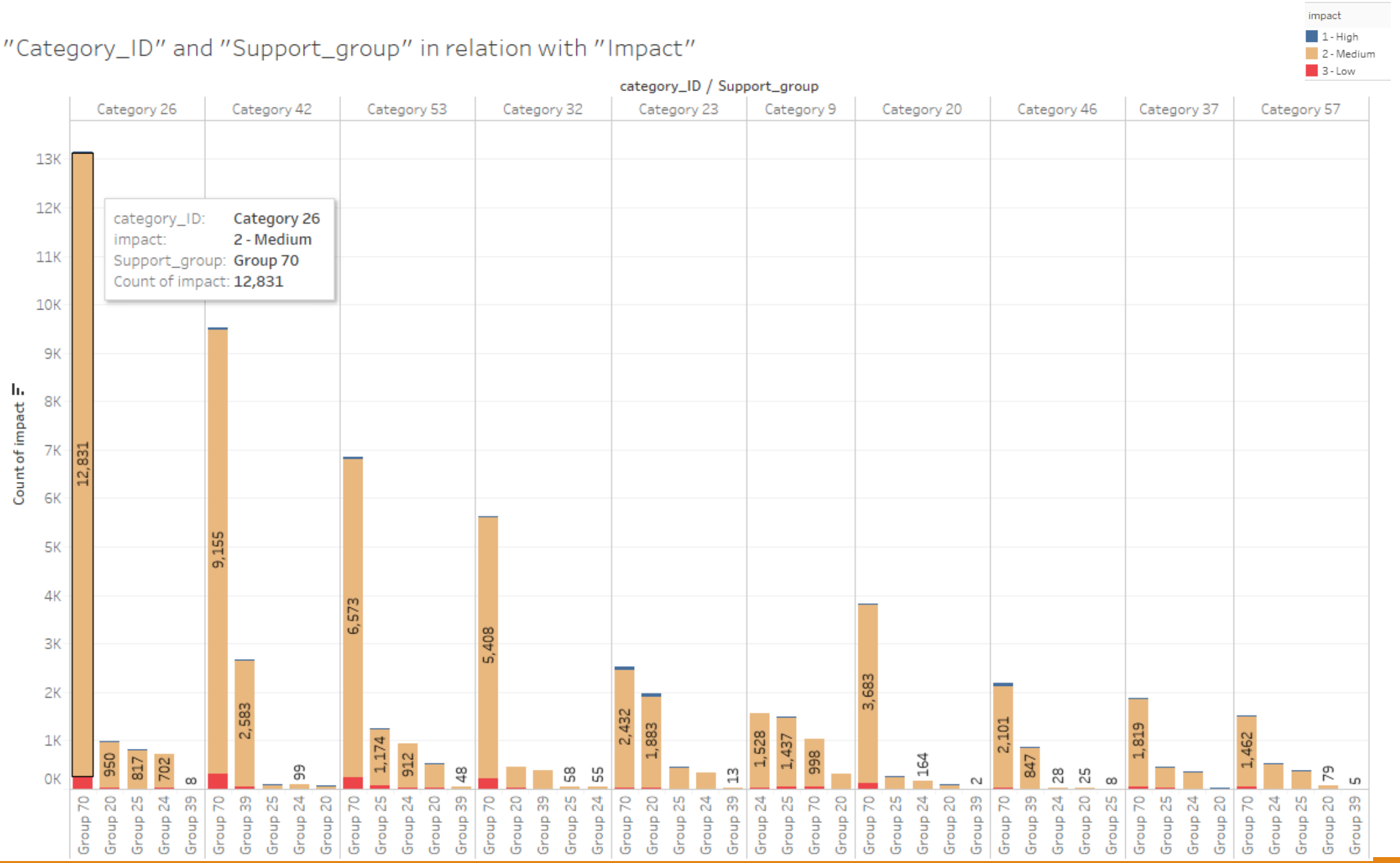


Certain particular "labels" from different attributes have a much higher effect in prediction for eg: category 26,42,53,46, location 204, 141, 108, 93, 51, active, new and resolved labels. **All 3 attributes have a say in prediction**

Impact of different labels of the predictors

Going through other attributes leads us to the following observations:

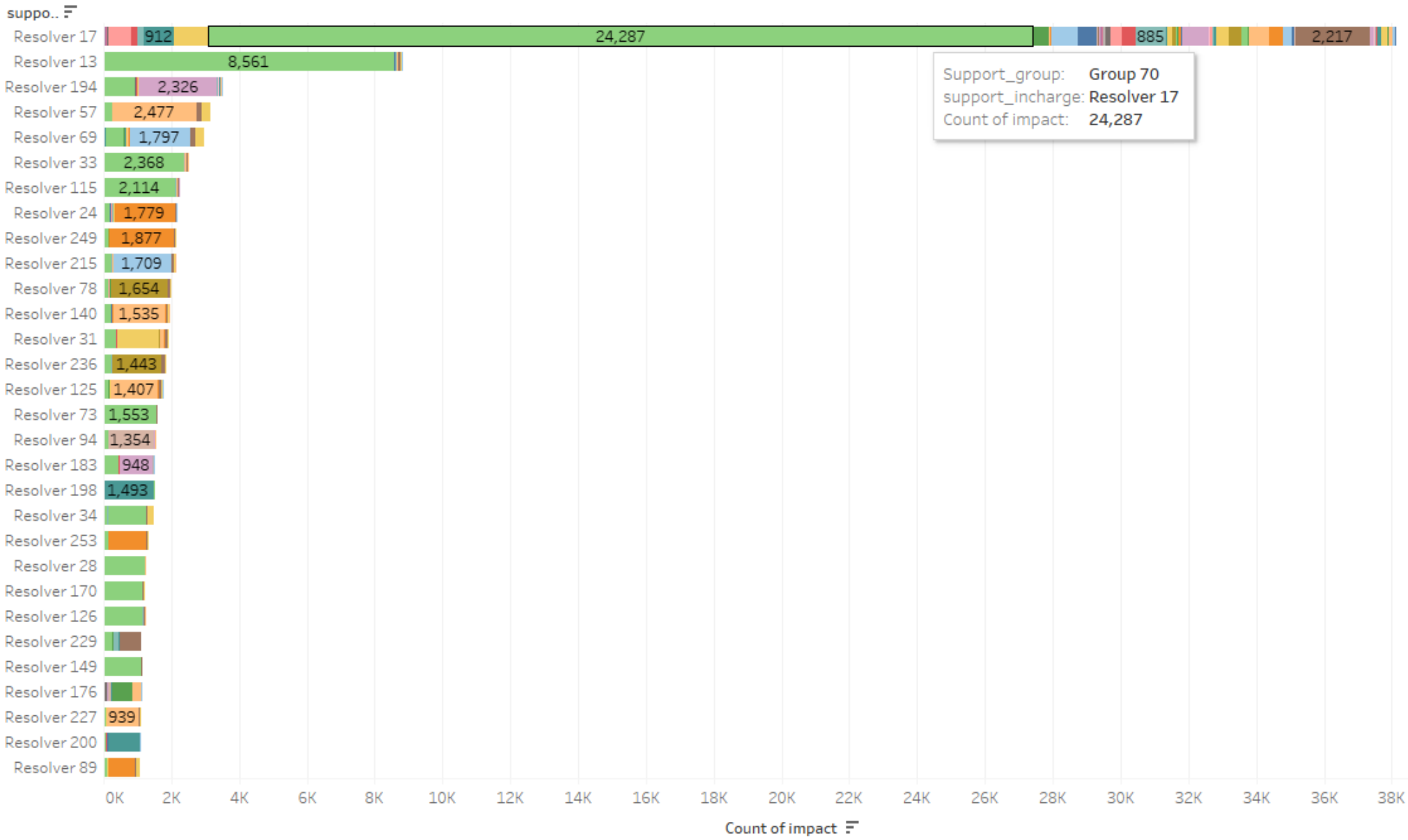
- Group 70 in "**support_grp**" account for 40.7%
- Sym 491 in "**user_symptom**" account for 59.9%
- Category 26,42,53,46,23 and 9 in "**category_id**" accounts for almost 50%
- Location 204,161,143 in "**location**" account for almost 57%
- Phone under "**type_contact**" account for 99%
- **“Updated_by”**908, 44, 60 account for 40%
- **“Created by”** 10 account for 55%
- True in **"confirmation check"** account for 71 %
- False in **"doc_knowledge"** account for 82%
- 4/7/16 and 17/3/16 have almost 37% values under **"created_at"**
- 0,1,2,3,4 in **“count_updated”** account for 64%
- 0 in **“count_opening”** account for 98%
- 0,1,2 in **“count_reassign”** account for almost 85%
- True in **"active"** account for 82%
- **“Change request”** and **“problem_id”** have almost 98% values missing



1. Certain Categories and Support groups have majority of the incidents
2. For eg: "Category_id" 26, and "Support_group" 70, 20, 25, 24, 39 have maximum incidents

"Category" and "Sub-category" have an effect on "Impact"

"Support_group" + "Support_incharge" and "Impact"

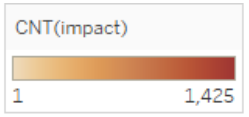


1. "Support_Incharge" (17) have dealt with majority of the incidents
2. Majority of the "Support_Incharge" who dealt with issues belong with similar "Support_group" (eg:70) irrespective of "Impact" classes

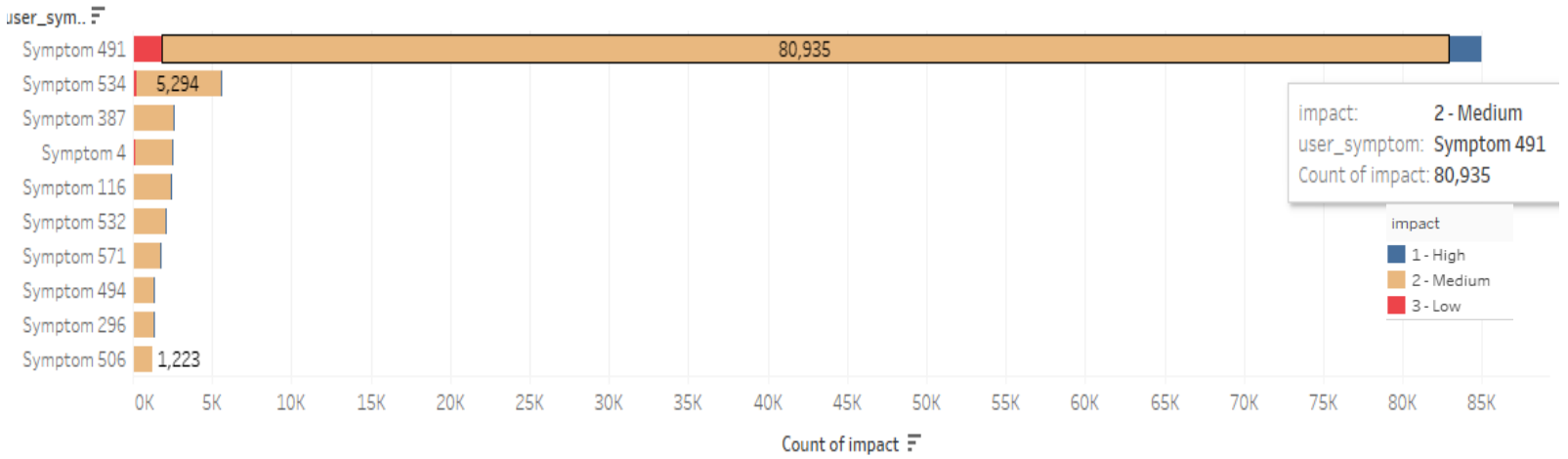
"Support_incharge" and "Support_group" doesnot have much prediction power

"ID_Caller" s effect on "Impact"

ID_cal..	impact: 1 - High	2 - Medium	3 - Low	Grand Total
Grand Total	3,491	1,34,335	3,886	1,41,712
Caller 1904	4	1,425	25	1,454
Caller 290	4	408	379	791
Caller 4514	5	711		716
Caller 1441	48	274		322
Caller 298		293		293
Caller 3763	33	229	8	270
Caller 93	39	194	6	239
Caller 1531		228	3	231
Caller 4414		217	7	224
Caller 3160		207	13	220
Caller 90		219		219
Caller 2471	16	198	5	219
Caller 3479		150	57	207
Caller 1270		207		207
Caller 363		204		204
Caller 3870	6	195		201
Caller 1517		195	3	198
Caller 707	20	176		196
Caller 5093		167	27	194
Caller 994		191		191
Caller 4180	5	122	60	187
Caller 5317		186		186
Caller 4808		186		186
Caller 156		186		186
Caller 501	7	171	7	185
Caller 2522		185		185
Caller 3038	11	148	25	184
Caller 2737		184		184
Caller 742		181		181



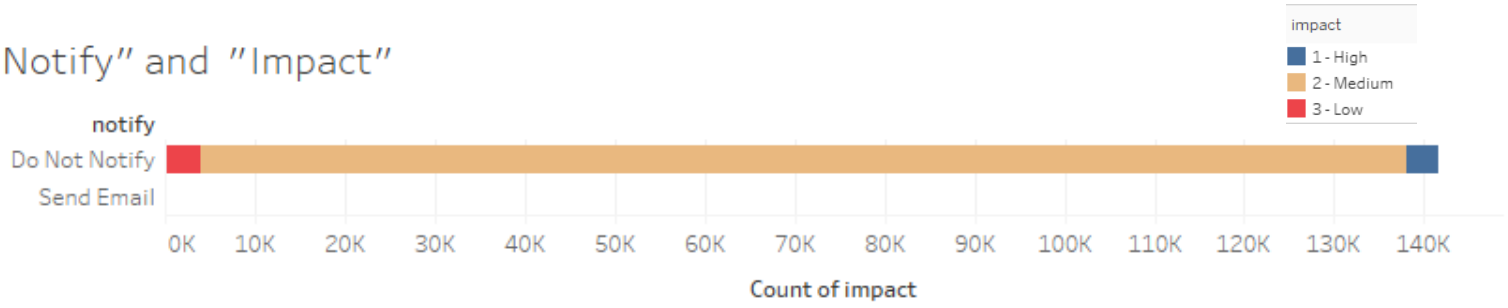
"User symptom" and "Impact"



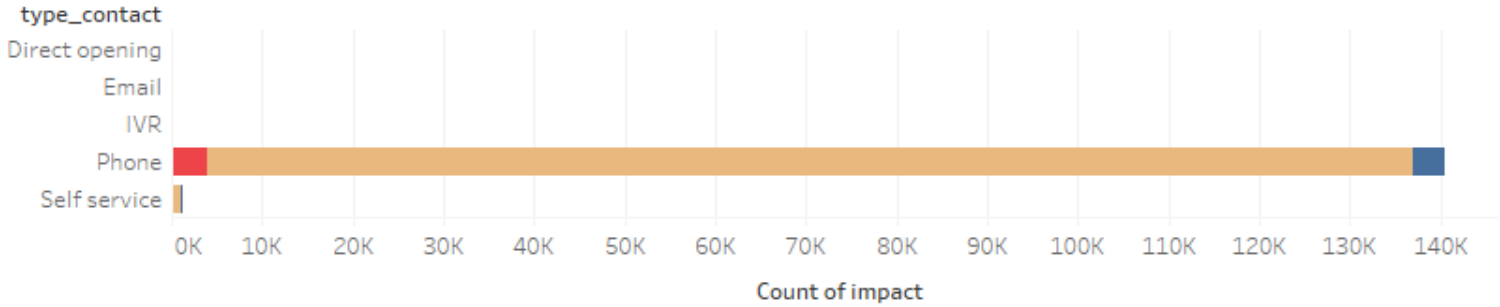
1. Certain "ID_Caller" have a high count of incidents, while the rest have much lower counts. There are majorly 4 cluster (>=700), (<700 and >=150), (<150 and >=50) and (<50)
2. "User symptoms" 491, and 534 seems to have majority of the incidents, irrespective of the incident type, rendering other values much powerless, with very minor incidents getting registered to others

"ID_caller" seem to have an effect on "Impact" while "User symptoms" doesn't seem to have an effect on the "Impact"

"Notify" and "Impact"



"Type_contact" and "Impact"



1."Notify", "Type_contact", "Active" have same distribution irrespective of "Impact" class

2."Confirmation_Check" and "Doc_knowledge" have almost equal proportion of "Impact" classes for both true and false criteria

None of the attributes have major role in prediction and can be discarded

"Confirmation_check" and "Impact"

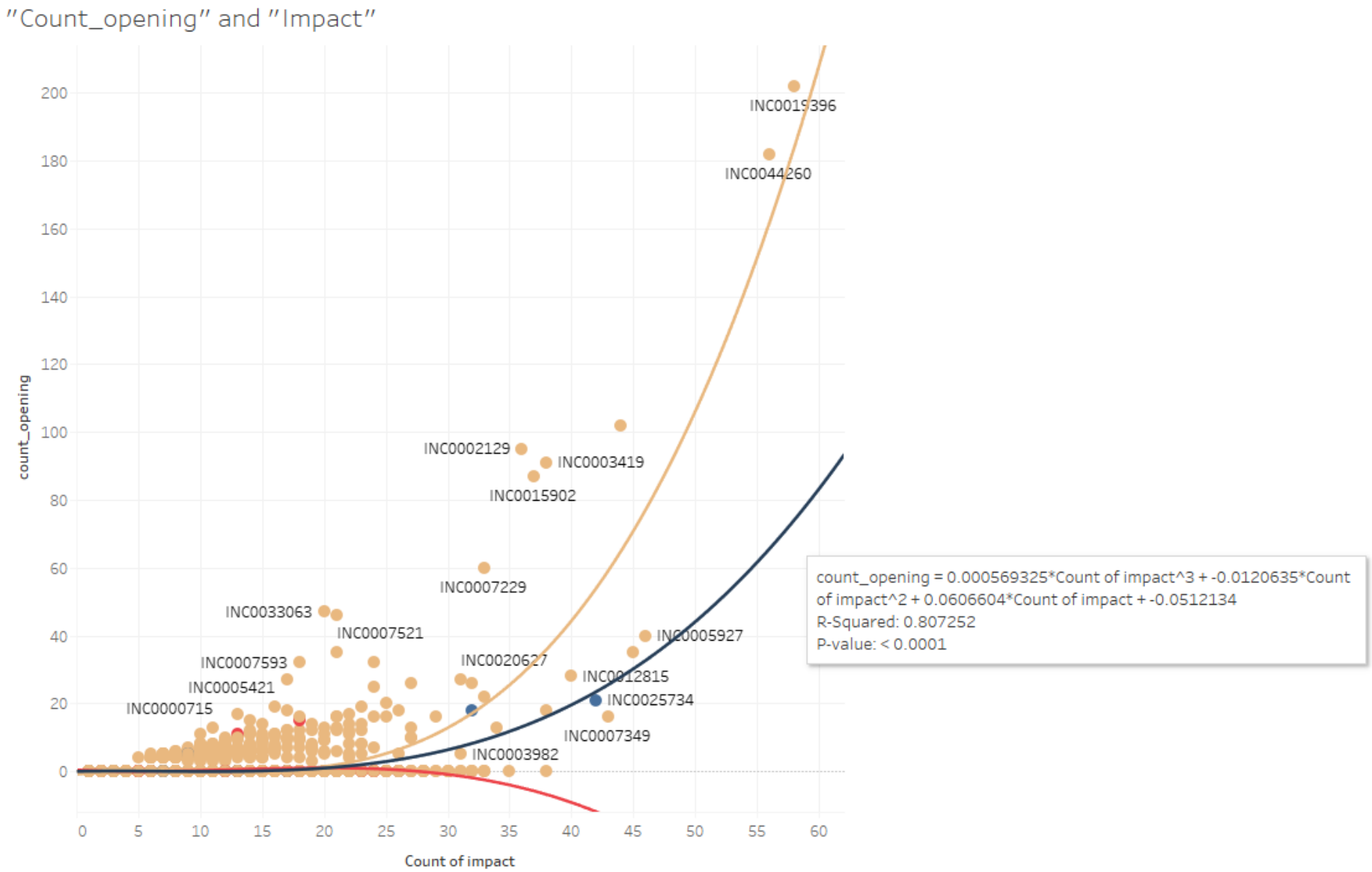
	impact		
confirmatio..	1 - High	2 - Medium	3 - Low
False	2.14%	94.92%	2.94%
True	3.25%	94.50%	2.25%

"Doc_knowledge" and "Impact"

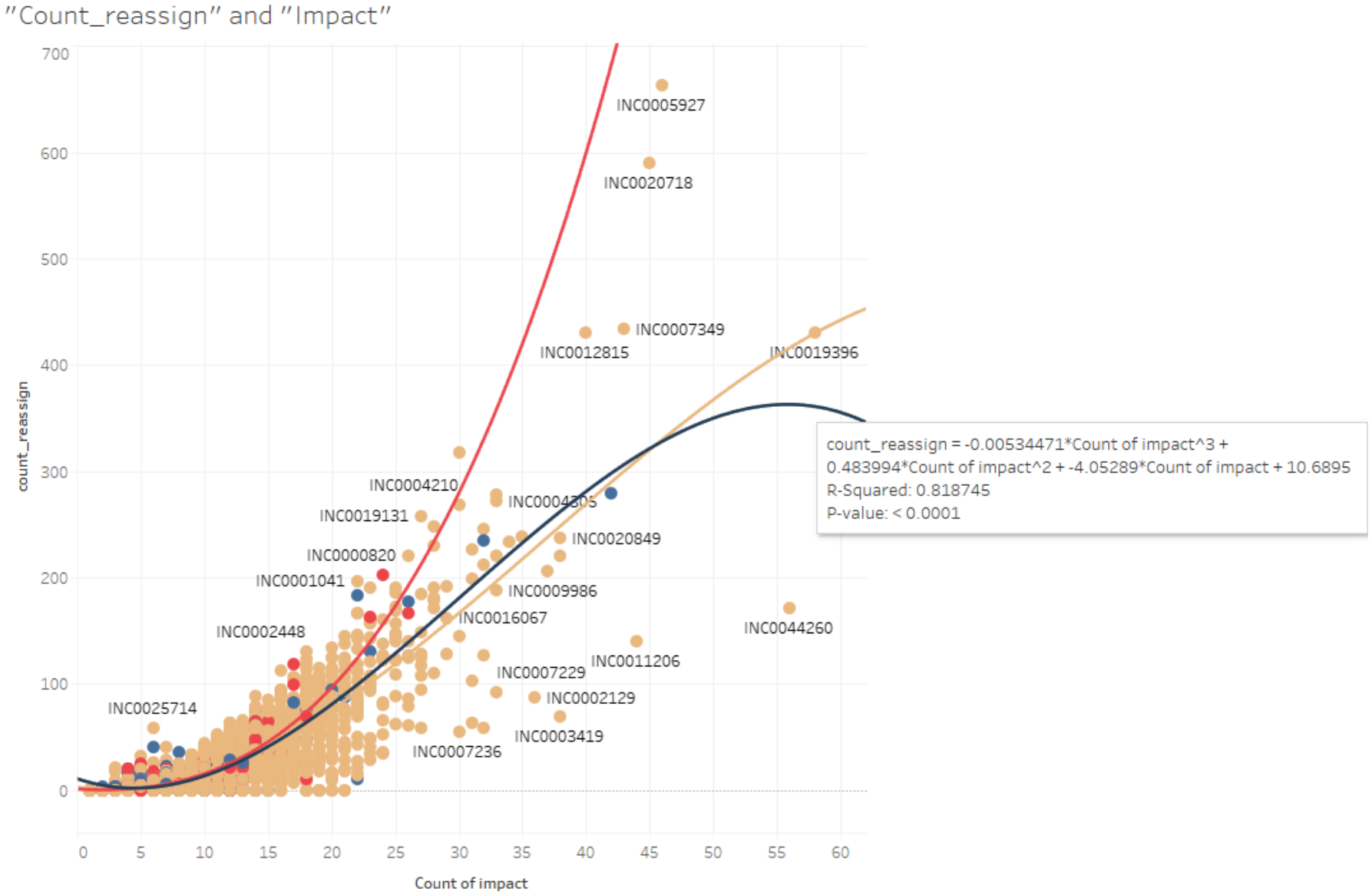
	impact		
Doc_knowl..	1 - High	2 - Medium	3 - Low
False	2.38%	95.59%	2.03%
True	2.85%	91.13%	6.02%

"Active" and "Impact"

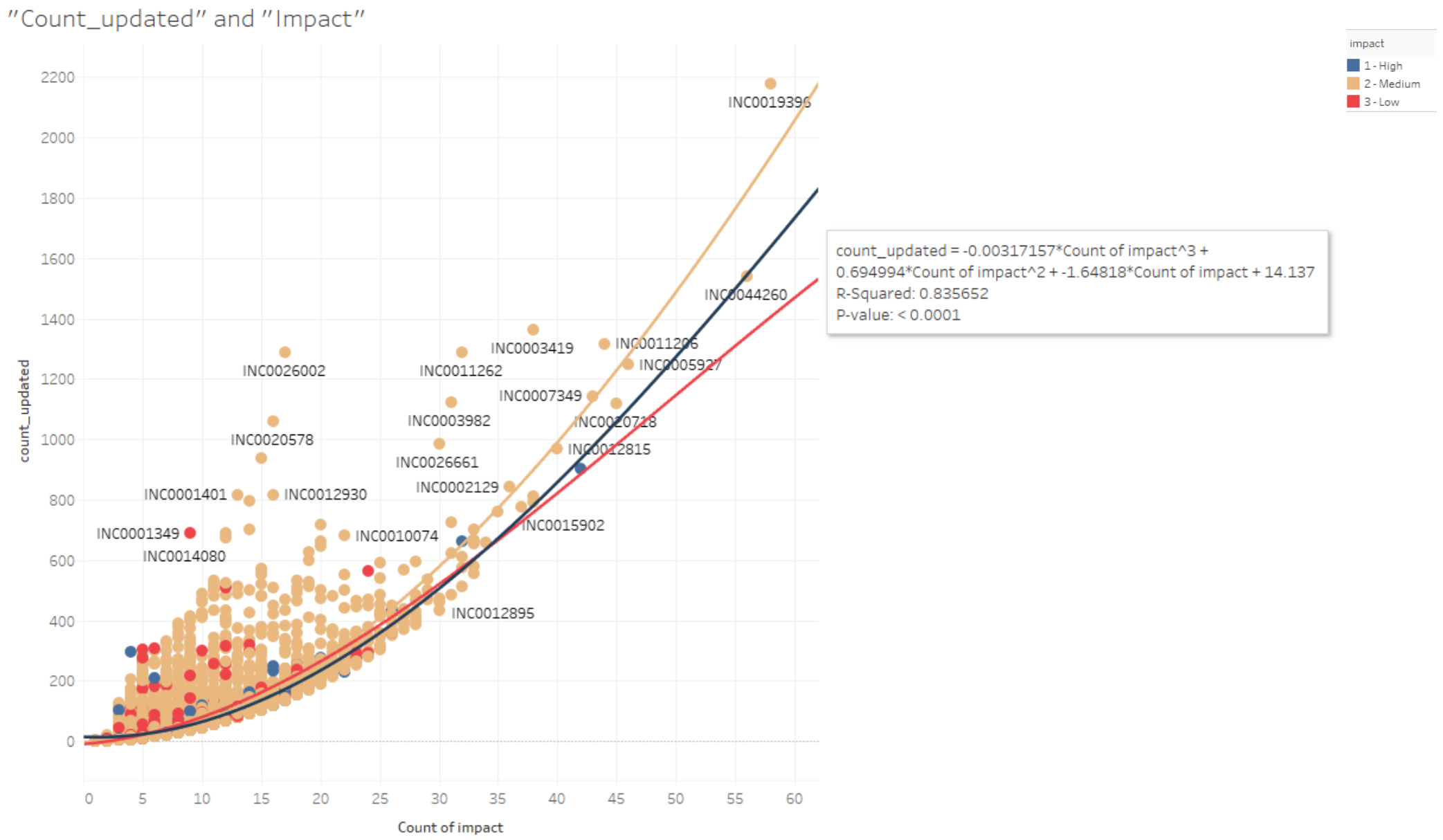
	impact		
active	1 - High	2 - Medium	3 - Low
False	1.69%	95.29%	3.02%
True	2.63%	94.69%	2.68%



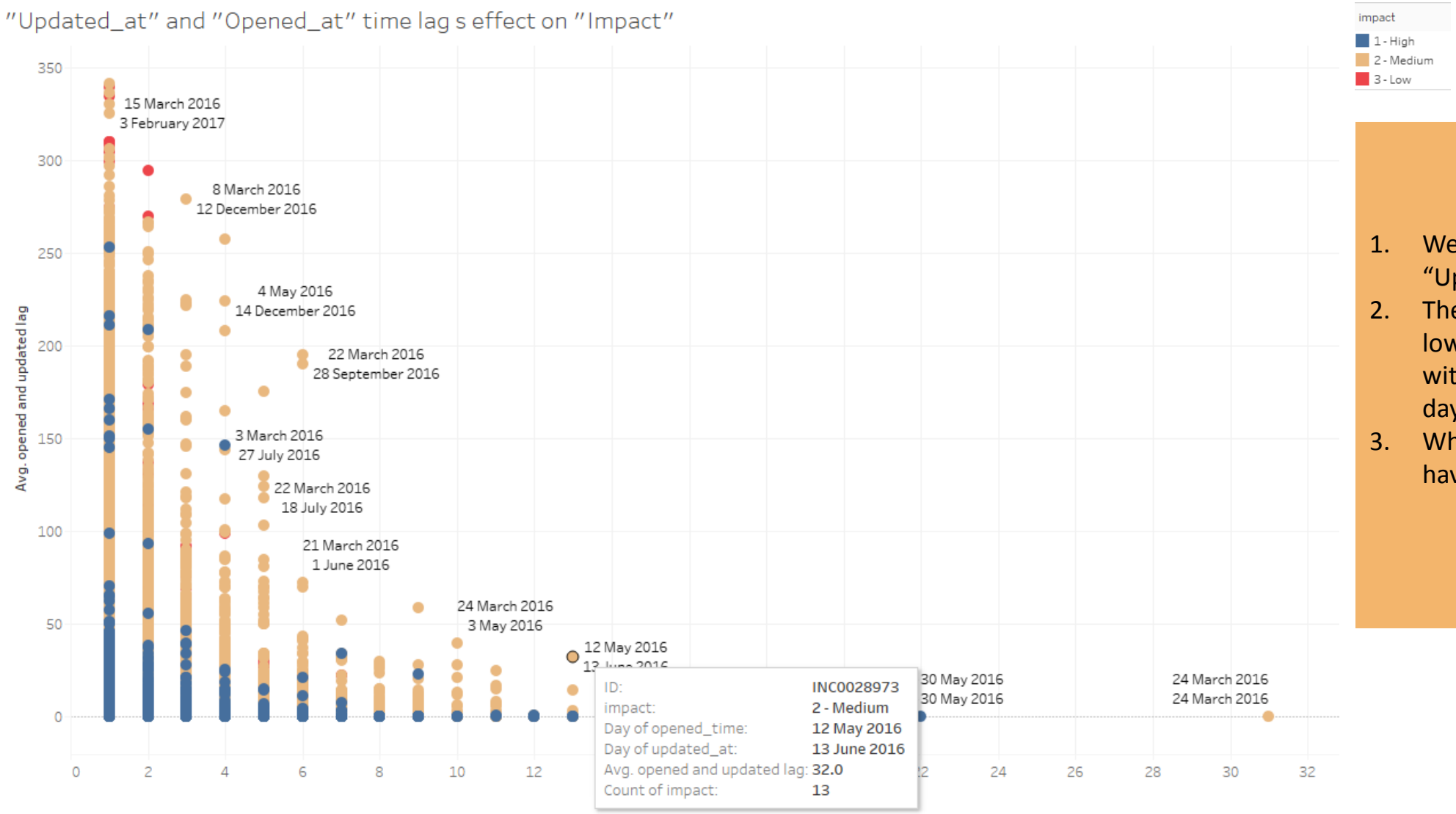
"Count_opening" and "High Impact" doesn't have a direct linear correlation but have a correlation value of 0.80, depicting positive correlation, using a 3rd degree polynomial model, majority of its value are "zero" irrespective of the impact class, so it can be discarded from prediction (rarely users have rejected the solution)



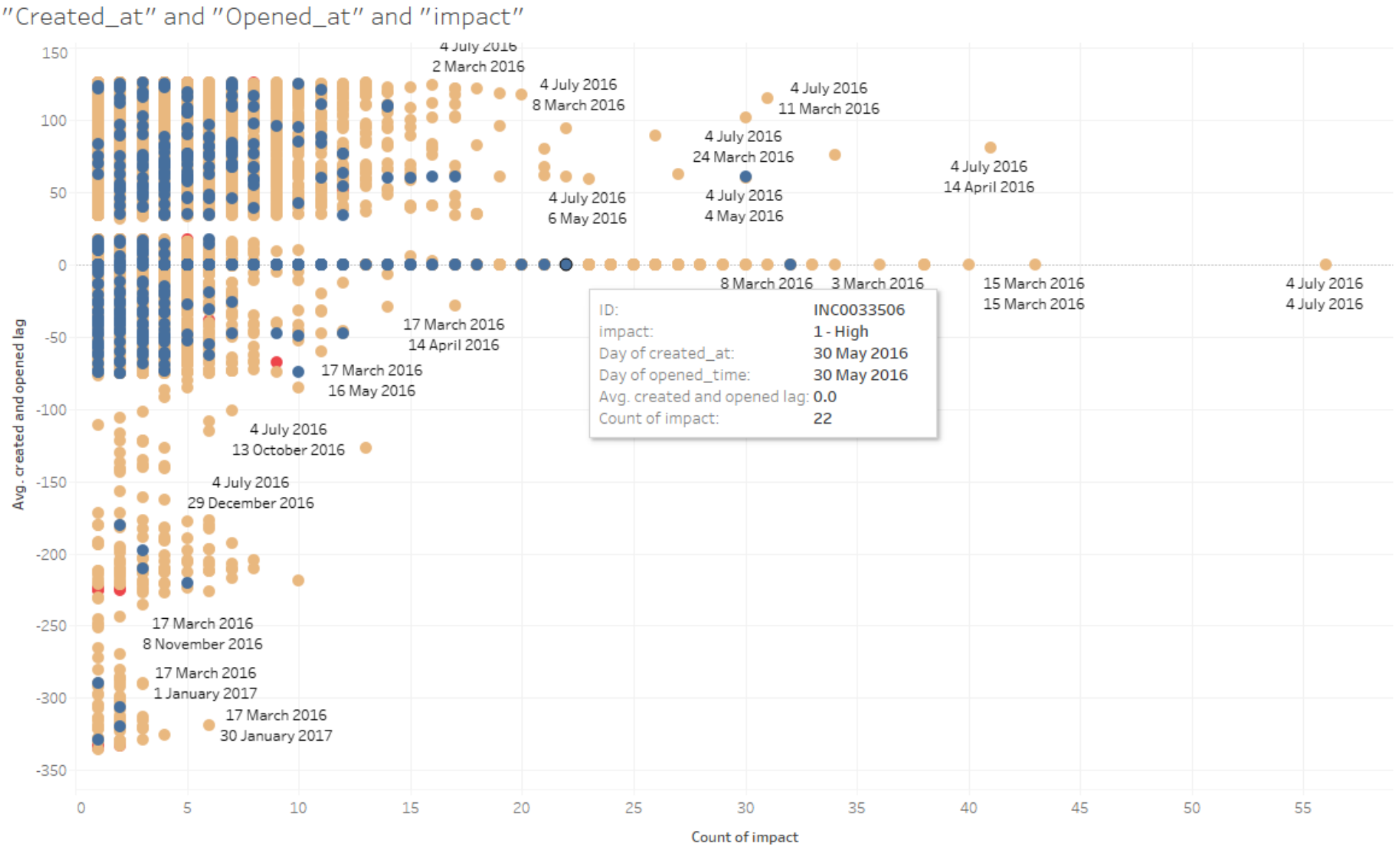
"Count_reassign" and "High Impact" have a correlation value of 0.818, depicting positive correlation, using a 3rd degree polynomial model(multiple number of times support group have been changed), majority datapoint from all classes fall in similar range



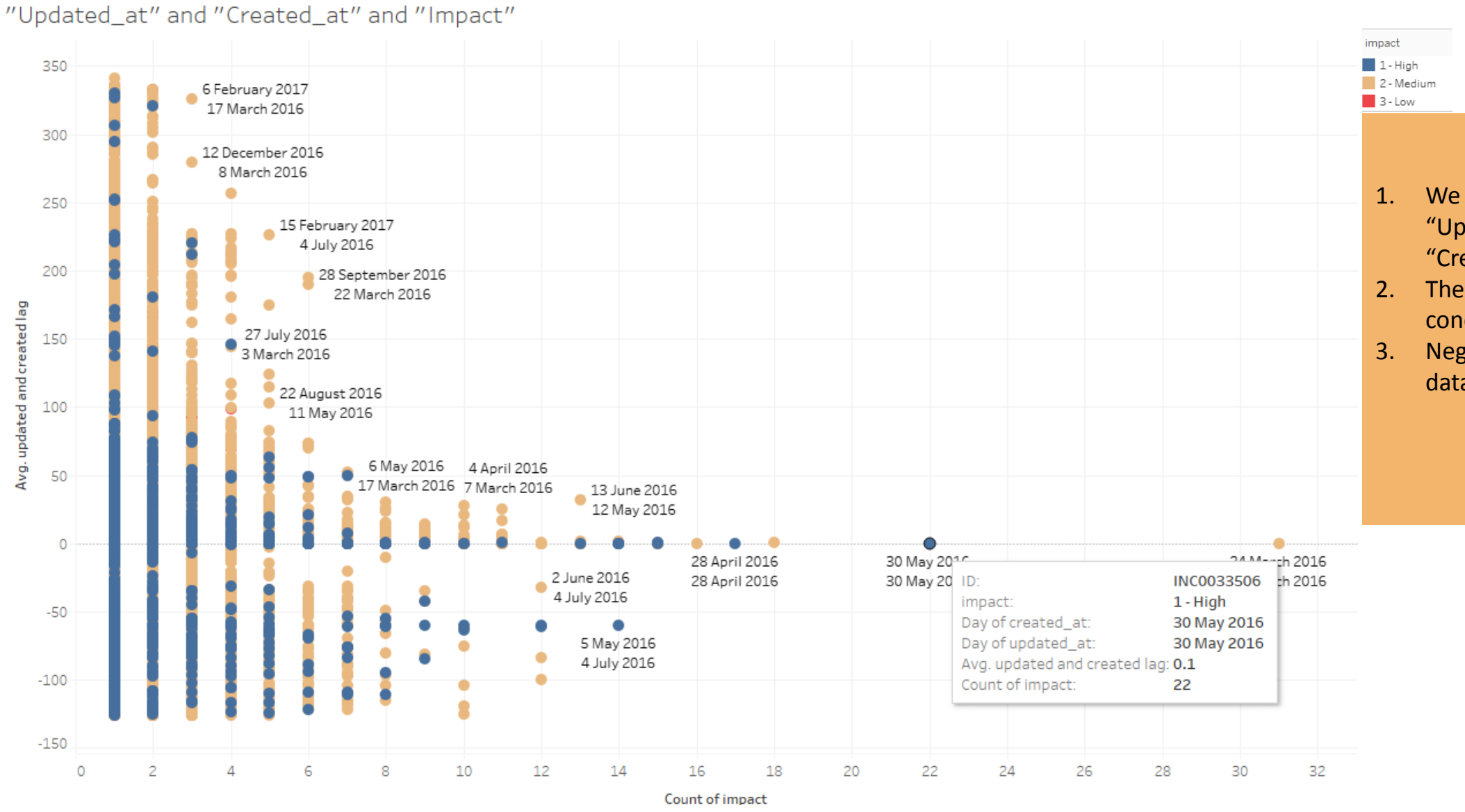
1. "Count_updated" and "High Impact" have a correlation value of 0.835, depicting positive correlation, using a 3rd degree polynomial model



1. We measured the "Opened_at" and "Updated_at" time lag in days
2. The "High Impact" incidents have a lower "opened and updated time lag", with majority in the range of 0-250 days, with average 0-30 days
3. While "Medium and Low Impact" have a varied range



1. We measured the “Created_at” and “Opened_at” time lag in days
2. Its visible that “High Impact” incidents are in the average range of (-75 to 125)
3. The spread is vast and is not conclusive for “Medium and Low Impact” incidents
4. Negative values depict some data entry anomaly



1. We measured the "Updated_at" and "Created_at" time lag in days
2. The spread is vast and is not conclusive
3. Negative values depict some data entry anomaly

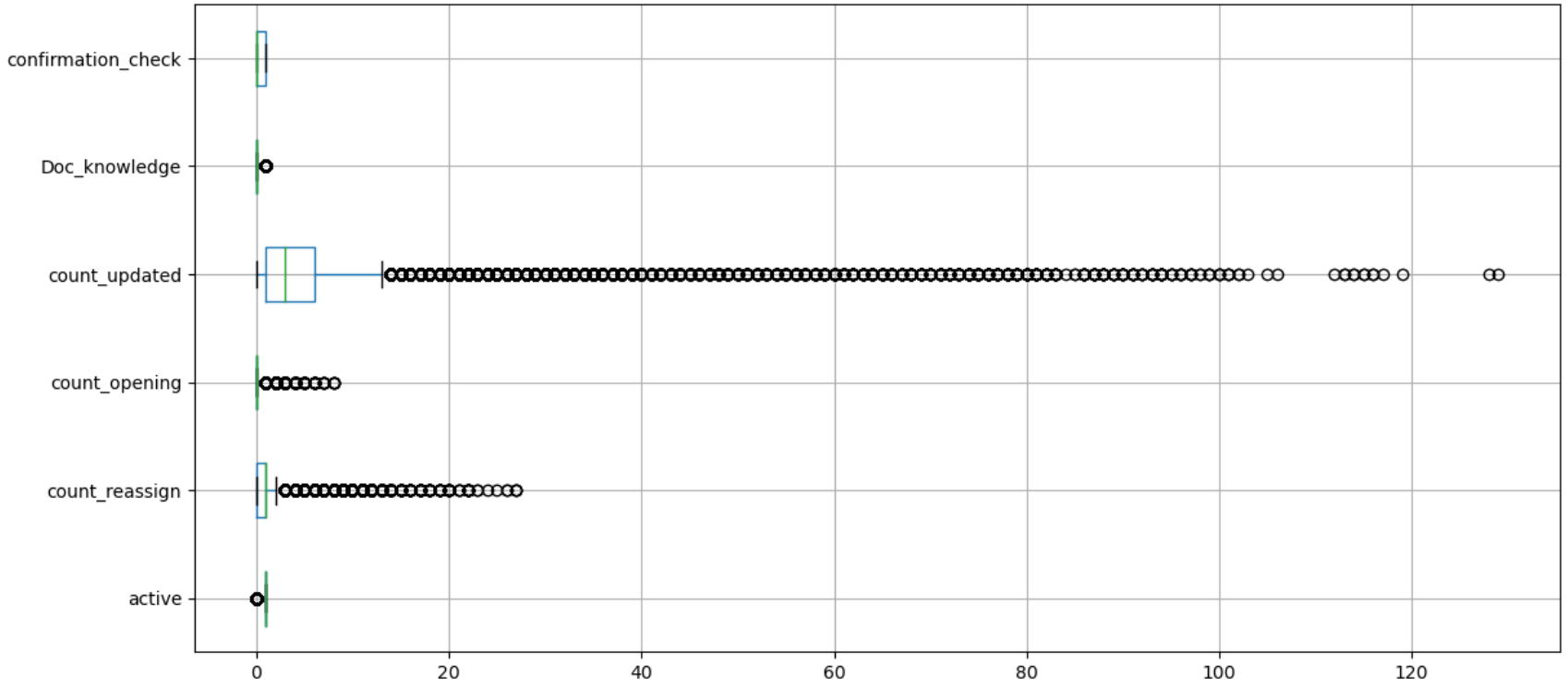
Important predictors

Attributes that might have a say in prediction as per visual understanding:

- “**Category”, “ID_status”, “Location”**
- **“Category”, “Sub_category”**
- **“ID_caller”**
- **“Count_updated”**
- **“Opened_at”, “Updated_at”**

While **“ count_opening”, “Change request”** and **“problem_id”** have almost 98% values missing and thus can be discarded, while rest attributes wont seem to have major prediction power.

Outliers and Categorical variables

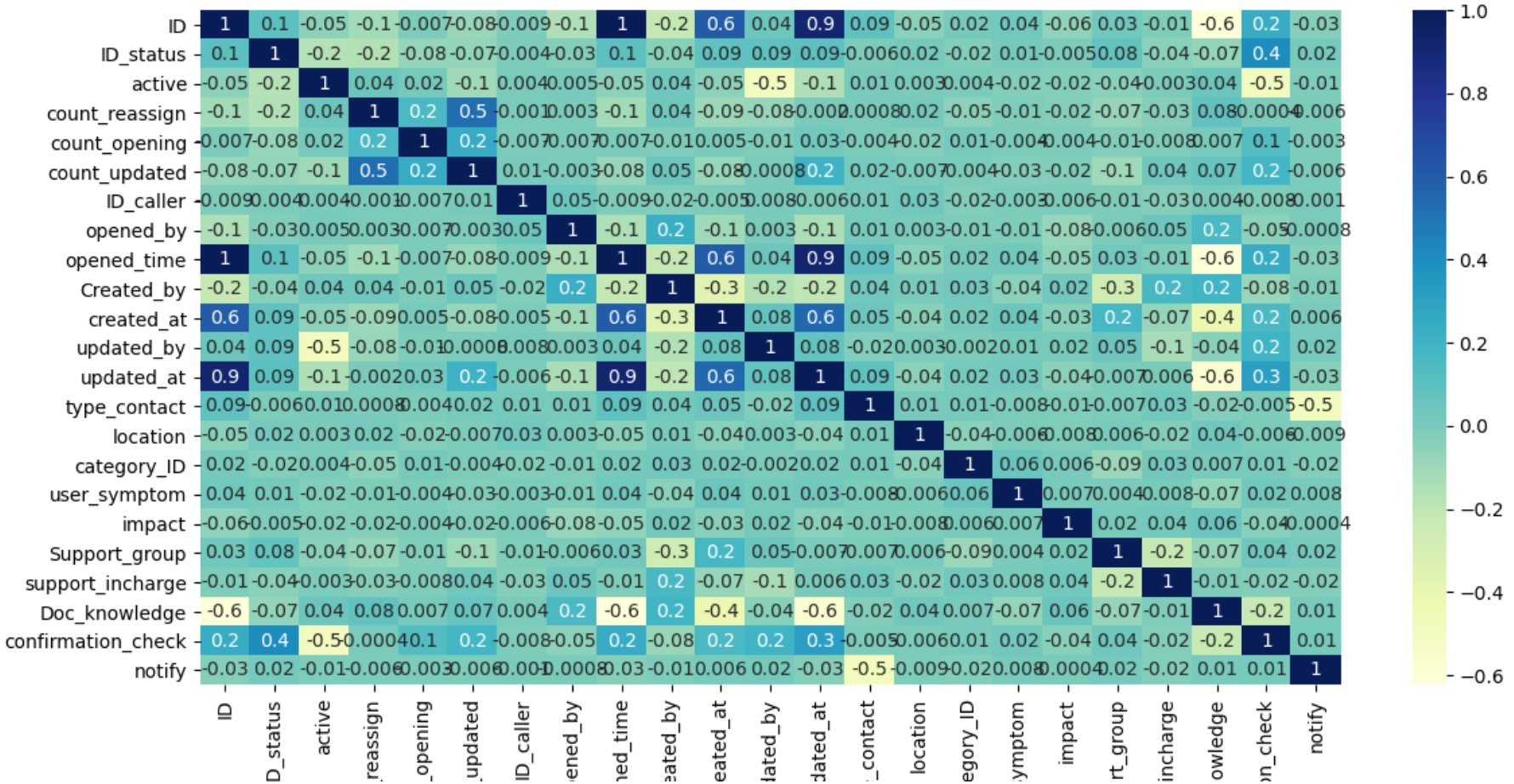


The int(), bool() have certain abnormal data points that might disturb the entire data distribution.

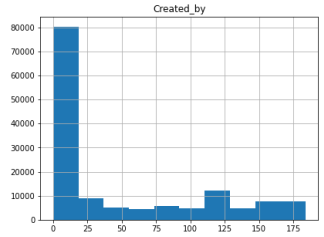
Particularly “count_updated” and “count_reassign” have major outliers

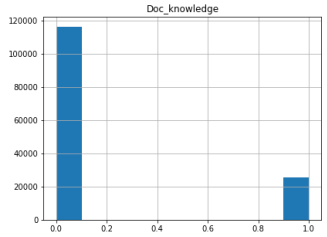
We also label encode categorical variables to prepare them for further usage

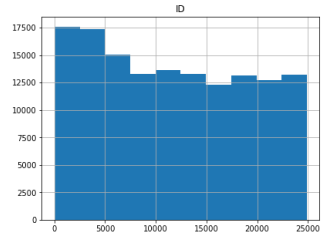
Establishing correlation between different attributes

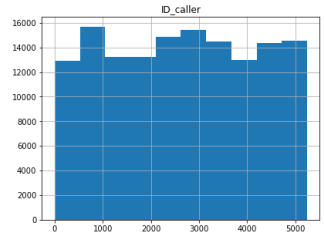


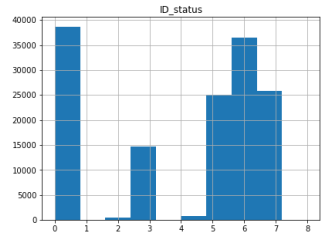
“opened_at” and “updated_at” have a correlation with “ID” and “opend_at” and “updated_at” are also related. Other attributes are not correlated to each other

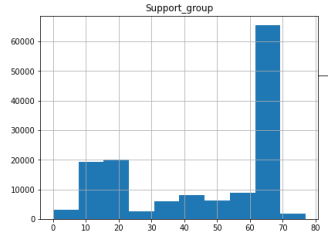


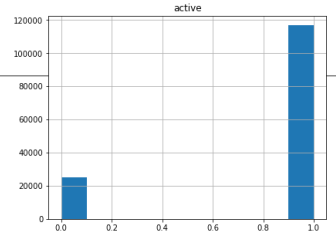


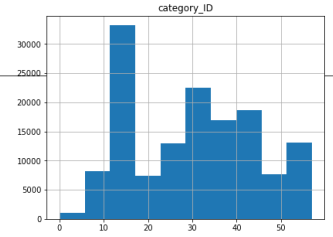


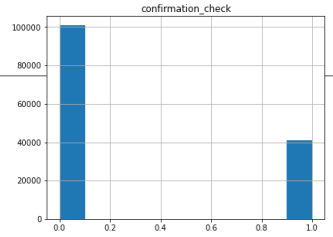


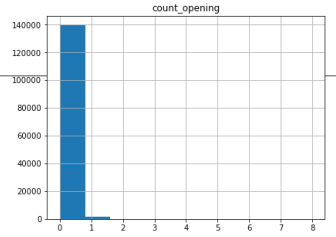


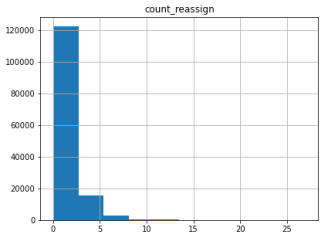


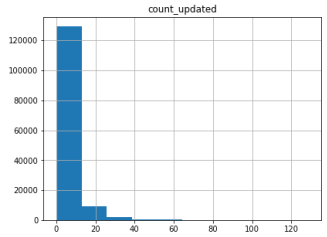


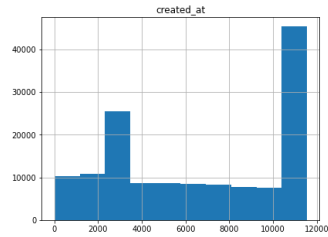


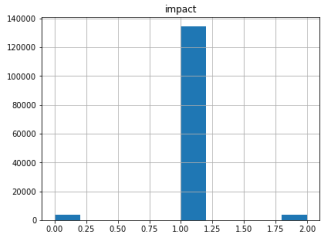


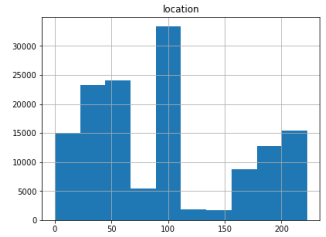


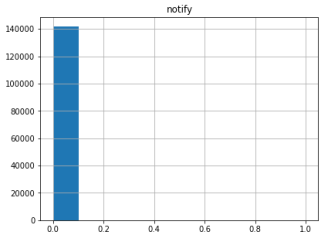


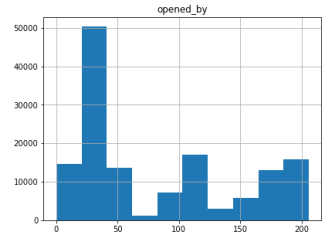


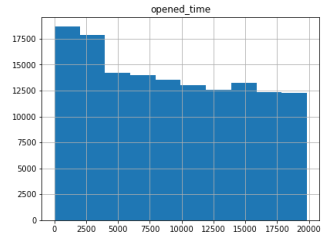


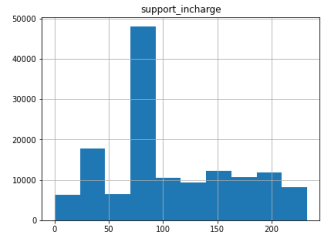


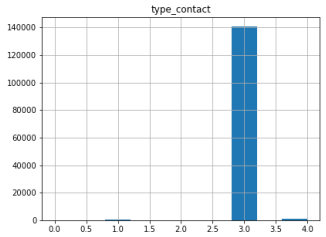


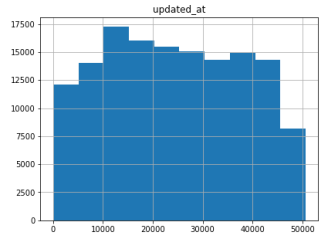


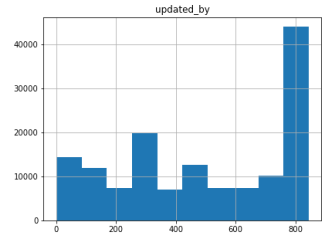


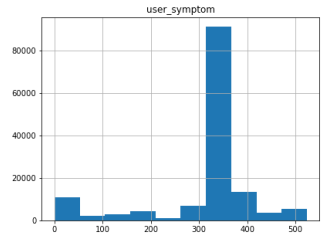












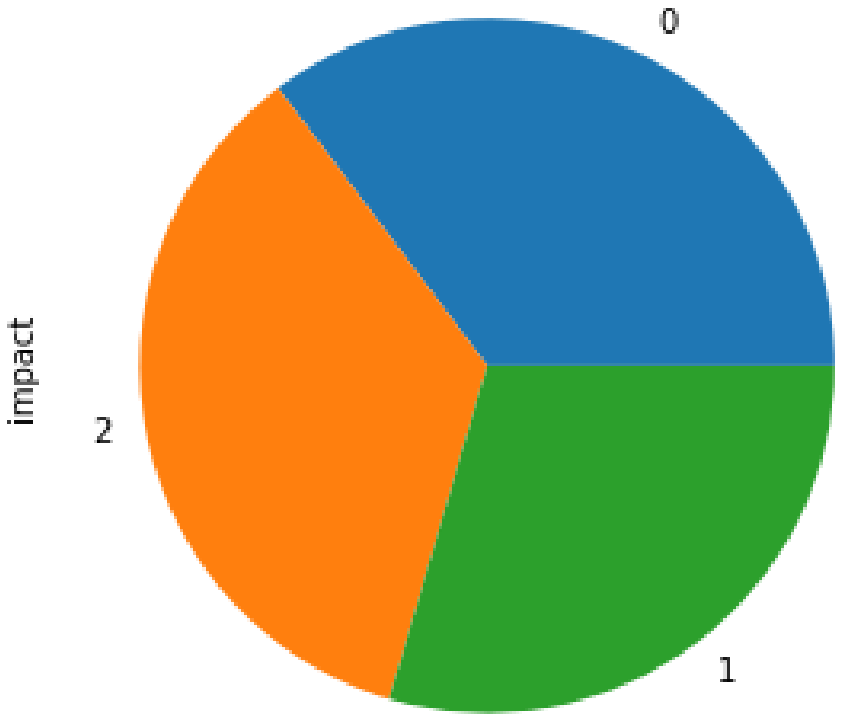
Normality test for each attribute

1.We did a qqplot() and a historgram plot to visualize the data symmetry.

2.Data distribution is not normal post label encoding. But still yielded better skewness and kurtosis values

3.We need to scale and standardize the data

Handle Imbalance



We tried to combine both SMOTE and undersampling as SMOTEENN, to generate a balanced dataset, where labels 0,1,2 denote “high”, “medium” and “low” incidents

Thank You