QMET 510 - Project Part 3 (FINAL)

Ankur Kumar – C00388588

Univariate analysis

List the variables of interest and give type(Numerical, categorical, ordinal)

Variables	Type
~~ID~~	~~Identification Number~~
~~name~~	~~label~~
~~category~~	~~Categorical~~
main_category	Categorical
~~currency~~	~~Categorical~~
~~deadline~~	~~Date~~
goal	Numerical
~~Launched~~	~~Date~~
pledged	Numerical
state	Categorical
backers	Numerical
~~country~~	~~Categorical~~
usd pledged	Numerical
usd_pledged_real	Numerical
usd_goal_real	Numerical
state_dummy_fail	Categorical
state_dummy_successful	Categorical

Screenshot of first few rows of data (BEFORE CLEAN UP)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	name	category	main_cate	currency	deadline	goal	launched	pledged	state	backers	country	usd pledg	usd_pledg	usd_goal_real
2	1000002330	The Songs	Poetry	Publishing	GBP	10/9/2015	1000	8/11/2015 12:12	0	failed	0	GB	0	0	1534
3	1000003930	Greeting F	Narrative	Film & Vic	USD	11/1/2017	30000	9/2/2017 4:43	2421	failed	15	US	100	2421	30000
4	1000004038	Where is I	Narrative	Film & Vic	USD	2/26/2013	45000	1/12/2013 0:20	220	failed	3	US	220	220	45000
5	1000007540	ToshiCapi	Music	Music	USD	4/16/2012	5000	3/17/2012 3:24	1	failed	1	US	1	1	5000
6	1000011046	Communi	Film & Vic	Film & Vic	USD	8/29/2015	19500	7/4/2015 8:35	1283	canceled	14	US	1283	1283	19500
7	1000014025	Monarch E	Restauran	Food	USD	4/1/2016	50000	2/26/2016 13:38	52375	successful	224	US	52375	52375	50000
8	1000023410	Support Sı	Food	Food	USD	12/21/2014	1000	12/1/2014 18:30	1205	successful	16	US	1205	1205	1000
9	1000030581	Chaser Str	Drinks	Food	USD	3/17/2016	25000	2/1/2016 20:05	453	failed	40	US	453	453	25000
10	1000034518	SPIN - Pre	Product D	Design	USD	5/29/2014	125000	4/24/2014 18:14	8233	canceled	58	US	8233	8233	125000
11	100004195	STUDIO IN	Document	Film & Vic	USD	8/10/2014	65000	7/11/2014 21:55	6241	canceled	43	US	6241	6241	65000
12	100004721	Of Jesus a	Nonfictior	Publishing	CAD	10/9/2013	2500	9/9/2013 18:19	0	failed	0	CA	0	0	2406
13	100005484	Lisa Lim N	Indie Rock	Music	USD	4/8/2013	12500	3/9/2013 6:42	12700	successful	100	US	12700	12700	12500
14	1000055792	The Cotta	Crafts	Crafts	USD	10/2/2014	5000	9/2/2014 17:11	0	failed	0	US	0	0	5000
15	1000056157	G-Spot Pla	Games	Games	USD	3/25/2016	200000	2/9/2016 23:01	0	failed	0	US	0	0	200000
16	1000057089	Tombston	Tabletop (	Games	GBP	5/3/2017	5000	4/5/2017 19:44	94175	successful	761	GB	57764	121857	6470
17	1000064368	Survival R	Design	Design	USD	2/28/2015	2500	1/29/2015 2:10	664	failed	11	US	664	664	2500
18	1000064918	The Beard	Comic Boo	Comics	USD	11/8/2014	1500	10/9/2014 22:27	395	failed	16	US	395	395	1500
19	1000068480	Notes Fro	Art Books	Publishing	USD	5/10/2015	3000	4/10/2015 21:20	789	failed	20	US	789	789	3000
20	1000070642	Mike Core	Music	Music	USD	8/17/2012	250	8/2/2012 14:11	250	successful	7	US	250	250	250

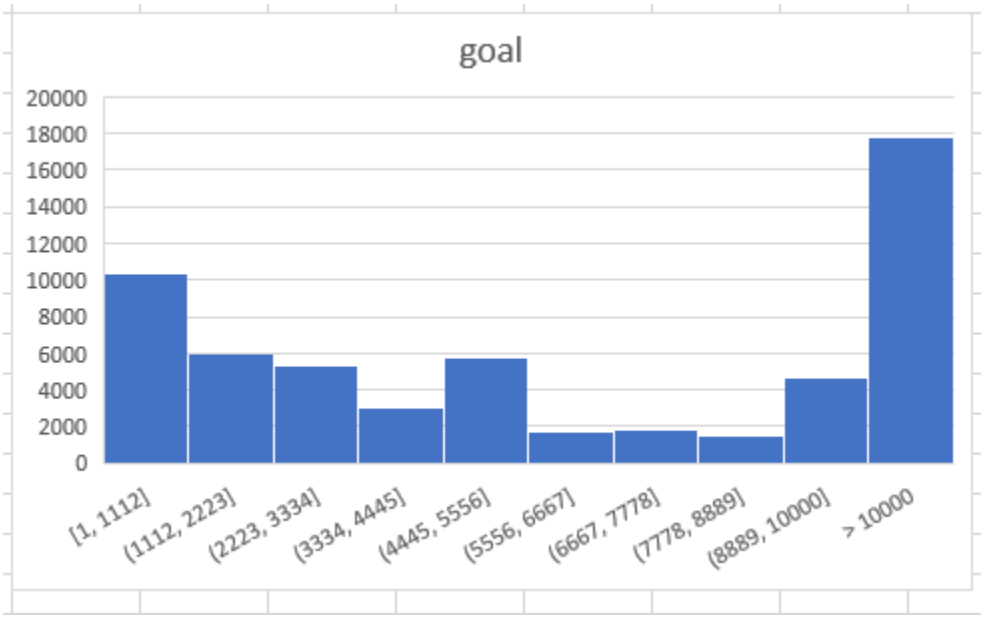
Screenshot of first few rows of data (AFTER CLEAN UP WITH DUMMY VARIABLES)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	ID	name	category	main_cate	currency	deadline	goal	launched	pledged	state	backers	country	usd pledg	usd_pledg	usd_goal_	state_dummy_fail	state_dummy_successful
2	1000002330	The Songs	Poetry	Publishing	GBP	10/9/2015	1000	8/11/2015	0	failed	0	GB	0	0	1534	1	0
3	1000003930	Greeting F	Narrative	Film & Vic	USD	11/1/2017	30000	9/2/2017	2421	failed	15	US	100	2421	30000	1	0
4	1000004038	Where is I	Narrative	Film & Vic	USD	2/26/2013	45000	1/12/2013	220	failed	3	US	220	220	45000	1	0
5	1000007540	ToshiCapi	Music	Music	USD	4/16/2012	5000	3/17/2012	1	failed	1	US	1	1	5000	1	0
6	1000014025	Monarch E	Restauran	Food	USD	4/1/2016	50000	2/26/2016	52375	successful	224	US	52375	52375	50000	0	1
7	1000023410	Support S	Food	Food	USD	12/21/2014	1000	12/1/2014	1205	successful	16	US	1205	1205	1000	0	1
8	1000030581	Chaser Str	Drinks	Food	USD	3/17/2016	25000	2/1/2016	453	failed	40	US	453	453	25000	1	0
9	100004721	Of Jesus a	Nonfictior	Publishing	CAD	10/9/2013	2500	9/9/2013	0	failed	0	CA	0	0	2406	1	0
10	100005484	Lisa Lim N	Indie Rock	Music	USD	4/8/2013	12500	3/9/2013	12700	successful	100	US	12700	12700	12500	0	1
11	1000055792	The Cotta	Crafts	Crafts	USD	10/2/2014	5000	9/2/2014	0	failed	0	US	0	0	5000	1	0
12	1000056157	G-Spot Pla	Games	Games	USD	3/25/2016	200000	2/9/2016	0	failed	0	US	0	0	200000	1	0
13	1000057089	Tombston	Tabletop (	Games	GBP	5/3/2017	5000	4/5/2017	94175	successful	761	GB	57764	121857	6470	0	1
14	1000064368	Survival R	Design	Design	USD	2/28/2015	2500	1/29/2015	664	failed	11	US	664	664	2500	1	0
15	1000064918	The Beard	Comic Boo	Comics	USD	11/8/2014	1500	10/9/2014	395	failed	16	US	395	395	1500	1	0
16	1000068480	Notes Fro	Art Books	Publishing	USD	5/10/2015	3000	4/10/2015	789	failed	20	US	789	789	3000	1	0
17	1000070642	Mike Core	Music	Music	USD	8/17/2012	250	8/2/2012	250	successful	7	US	250	250	250	0	1
18	1000071625	Boco Tea	Food	Food	USD	6/2/2012	5000	5/3/2012	1781	failed	40	US	1781	1781	5000	1	0
19	1000072011	CMUK. Sh	Fashion	Fashion	USD	12/30/2013	20000	11/25/2013	34268	successful	624	US	34268	34268	20000	0	1
20	1000081649	MikeyJ clc	Childrens	Fashion	AUD	9/7/2017	2500	8/8/2017	1	failed	1	AU	0	1	2026	1	0

DATA DISCUSSION

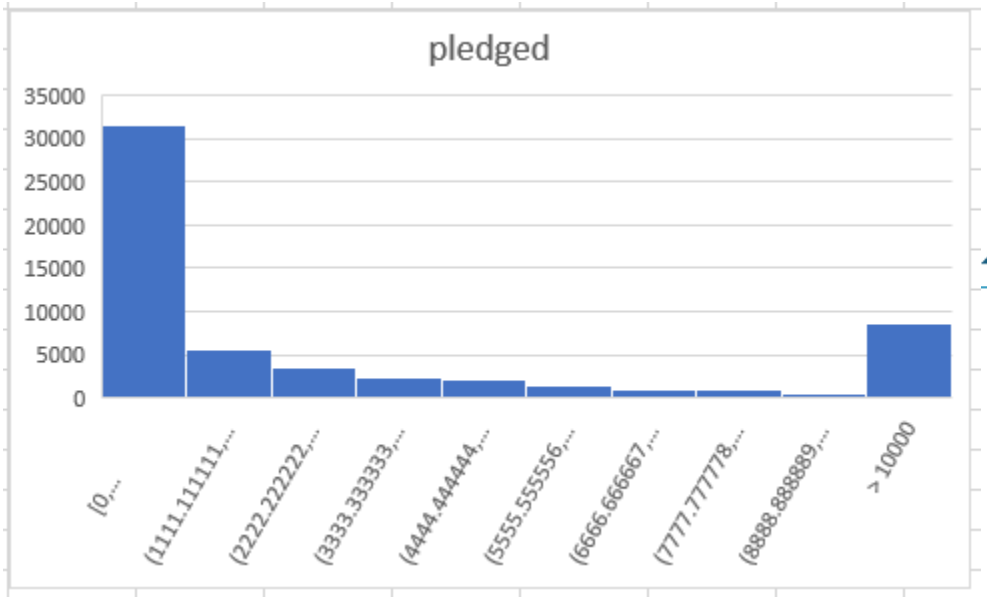
- The Data had a lot of missing values and had over 300,000 records. I had to trim the data because the excel was not able to handle operations for the original number of rows. Currently, the file contains 65,322 records.
- The values in various numerical variables had been missing, it has been handled by assigning them a value of ‘0’ because it was the smallest value which could be entered. I decided to use 0 after comparing it with other cell values.
- ID, deadline, launched, name, category column currency and country would be eliminated from the project because it is not required.
- States were originally divided into 5 categories :
    I. Failed
    II. Successful
    III. Canceled
    IV. Live
    V. Suspended
- After analyzing, I have decided to eliminate 3 of the state categories so I could filter out the result from **Failed** and **Successful**
- The current number of rows: **57392**
- I have also introduced 2 dummy variables: **state_dummy_fail, state_dummy_successful**

HISTOGRAMS AND SUMMER STATISTICS AFTER DATA CLEAN UP



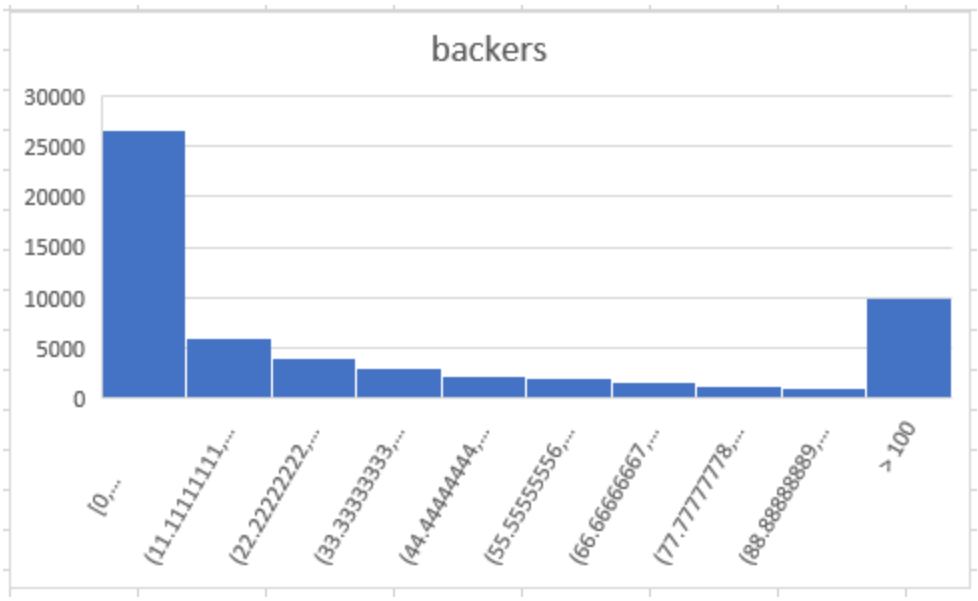
Statistics

Mean	17666.5063
Median	5000
Min	1
Max	550200
Standard Deviation	42804.1125
Unique Values	2371
Missing Values	0
Feature Type	Numeric Feature



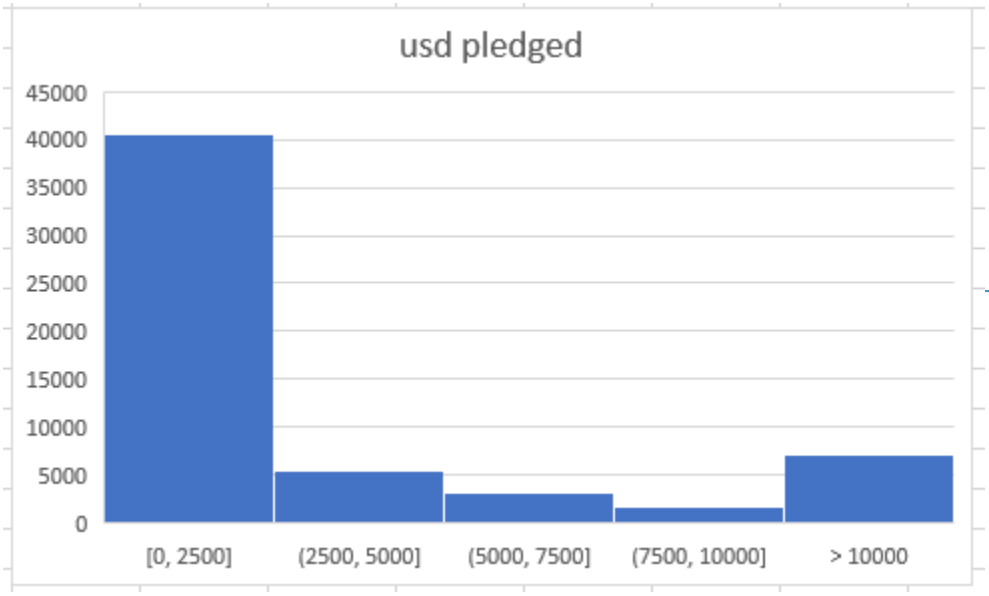
Statistics

Mean	10285.4403
Median	802
Min	0
Max	5545992
Standard Deviation	79059.1517
Unique Values	14904
Missing Values	0
Feature Type	Numeric Feature



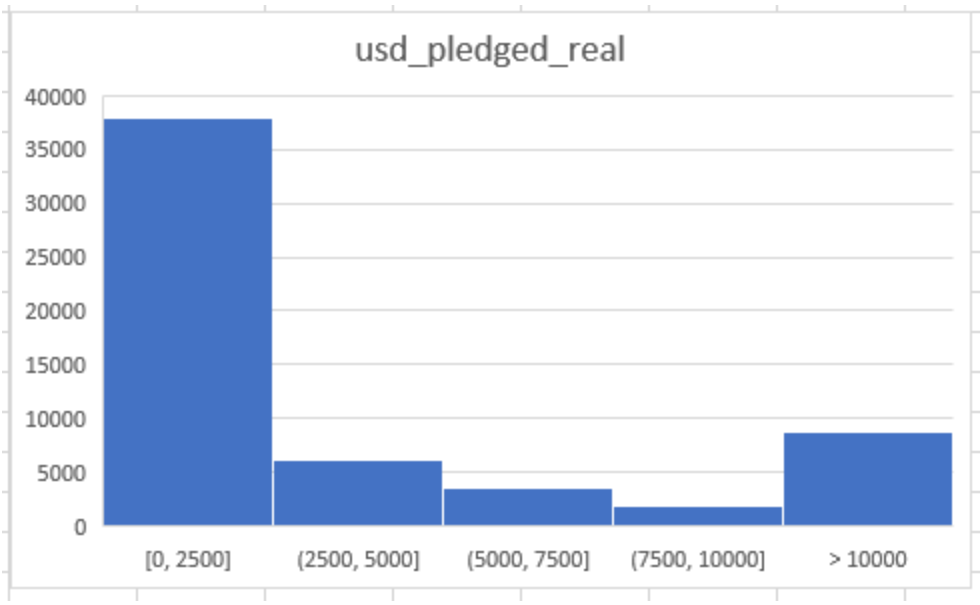
Statistics

Mean	113.226
Median	15
Min	0
Max	85581
Standard Deviation	788.2654
Unique Values	1789
Missing Values	0
Feature Type	Numeric Feature



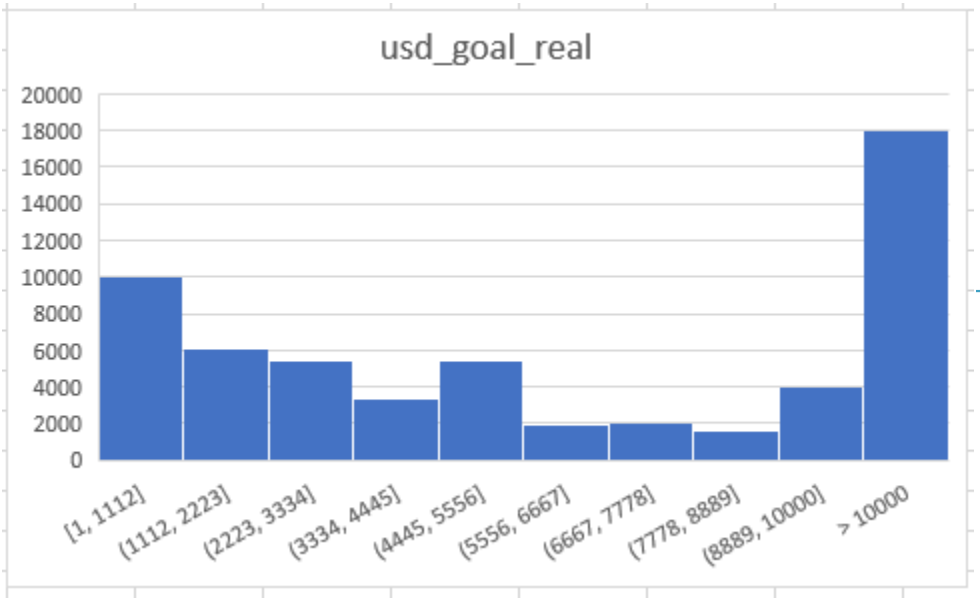
Statistics

Mean	7485.2747
Median	516
Min	0
Max	5545992
Standard Deviation	59674.5125
Unique Values	13415
Missing Values	0
Feature Type	Numeric Feature



Statistics

Mean	9727.9079
Median	804
Min	0
Max	5545992
Standard Deviation	74290.2092
Unique Values	14930
Missing Values	0
Feature Type	Numeric Feature



Statistics

Mean	16996.2536
Median	5000
Min	1
Max	778913
Standard Deviation	41354.0163
Unique Values	9141
Missing Values	0
Feature Type	Numeric Feature

OUTCOME VARIABLE - ***state***

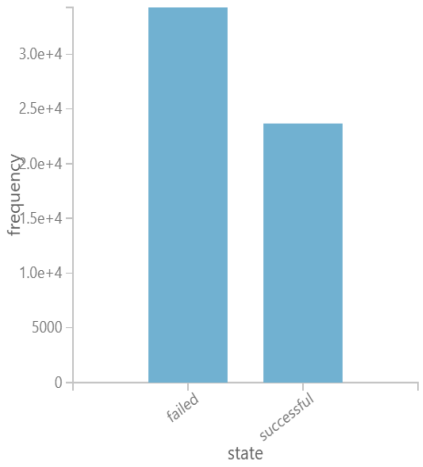
Statistics

Unique Values	2
Missing Values	0
Feature Type	String Feature

Visualizations

state

Histogram



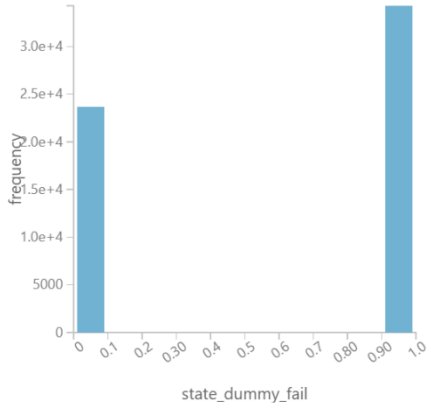
Statistics

Mean	0.5915
Median	1
Min	0
Max	1
Standard Deviation	0.4916
Unique Values	2
Missing Values	0
Feature Type	Numeric Feature

Visualizations

state_dummy_fail

Histogram



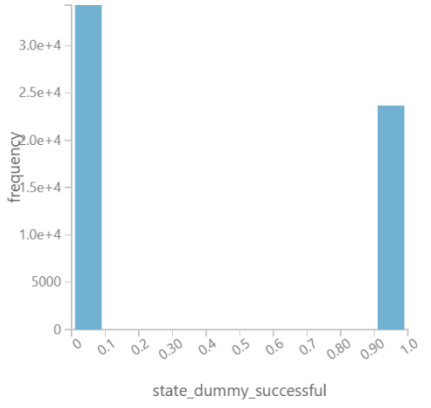
Statistics

Mean	0.4085
Median	0
Min	0
Max	1
Standard Deviation	0.4916
Unique Values	2
Missing Values	0
Feature Type	Numeric Feature

Visualizations

state_dummy_successful

Histogram



VISUALIZATION - ***main_category***

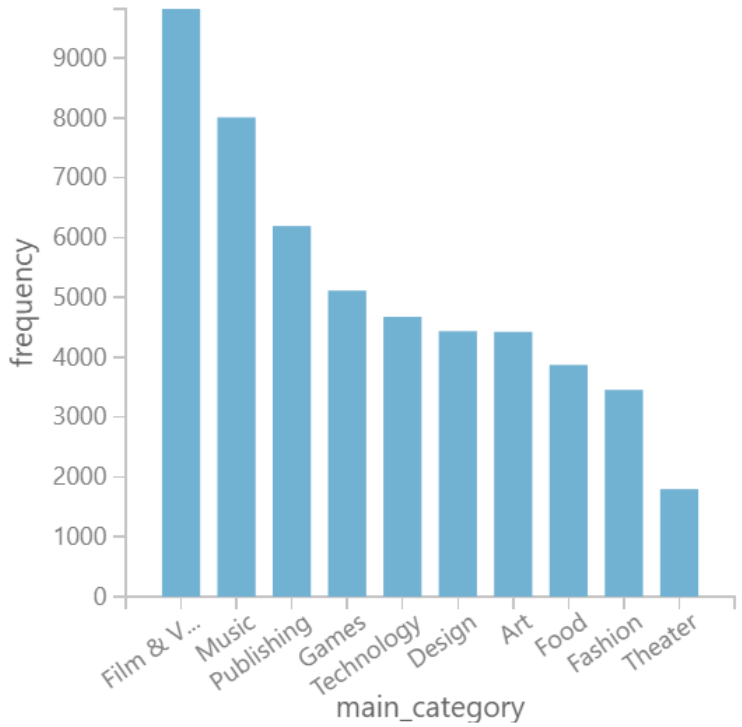
Statistics

Unique Values	15
Missing Values	0
Feature Type	String Feature

Visualizations

main_category

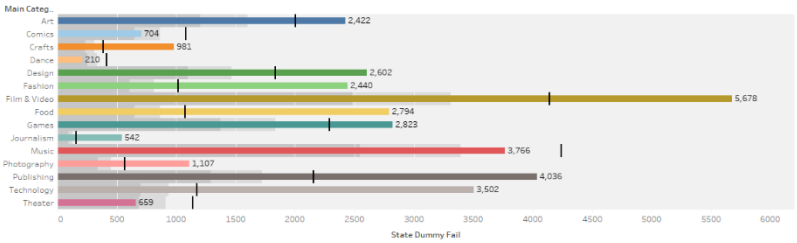
Histogram



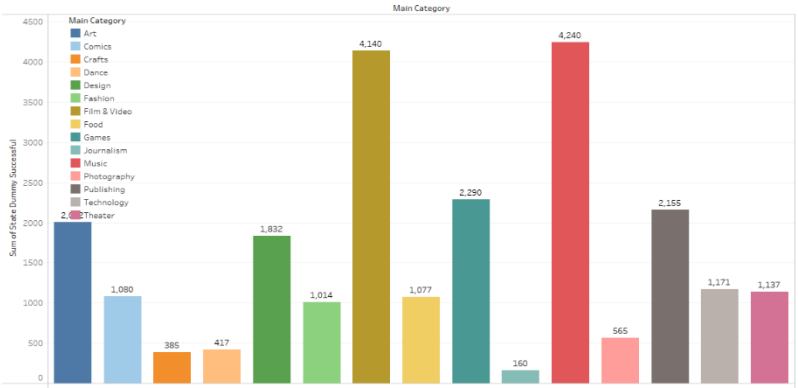
Bivariate Analysis

Tableau - Horizontal, side by side and Bullet graph showing the Fail and Successful state of main categories.

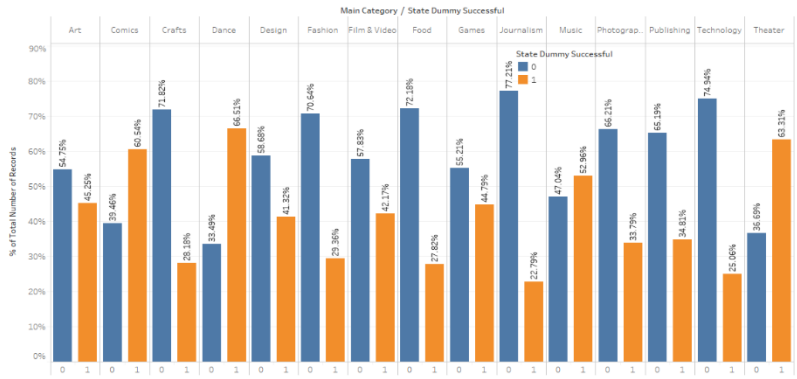
Horizontal bars

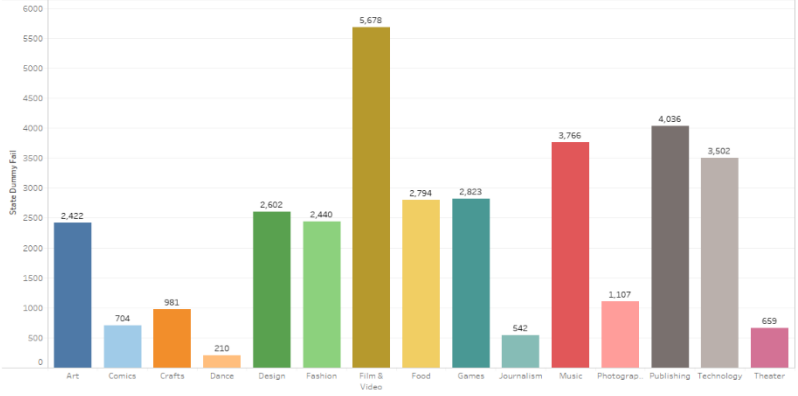


Fail vs Successful



Side by Side Bars





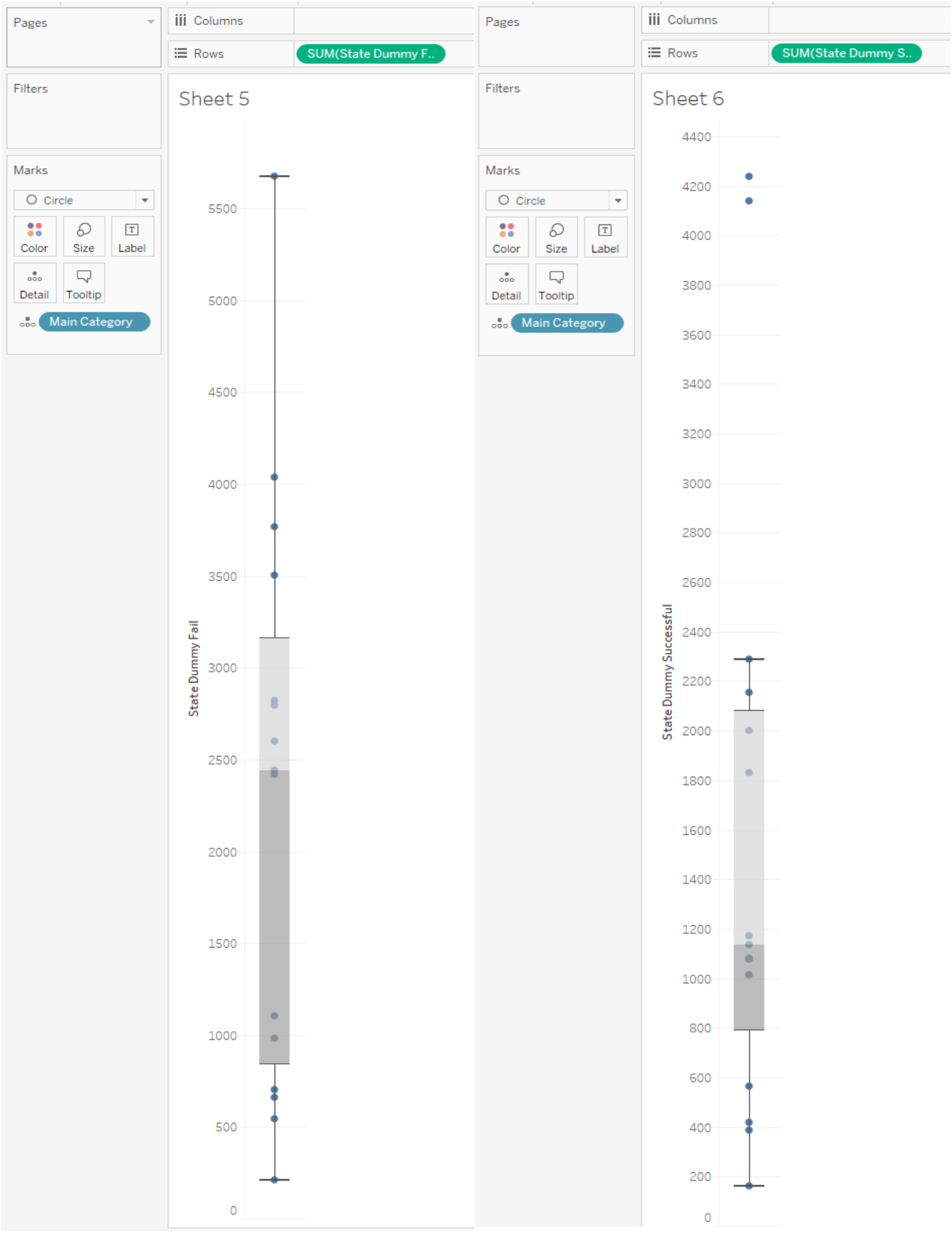
View Data: Sheet 4

Show aliases | Copy | Export All

Main Category	State Dummy Fail	State Dummy Successful
Theater	659	1,137
Technology	3,502	1,171
Publishing	4,036	2,155
Photography	1,107	565
Music	3,766	4,240
Journalism	542	160
Games	2,823	2,290
Food	2,794	1,077
Film & Video	5,678	4,140
Fashion	2,440	1,014
Design	2,602	1,832
Dance	210	417
Crafts	981	385
Comics	704	1,080
Art	2,422	2,002

Summary | Full Data | 15 rows

Data sheet showing various main categories and their following state in numbers.



Median for Fail state: 2440

Median for Success state: 1137

Microsoft Excel - Pivot tables of Categorical Variables and Outcome variable

main_category	Count of main_category	Percent	Average of state_dummy_fail	Average of state_dummy_successful
Art	4424	7.64%	54.75%	45.25%
Comics	1784	3.08%	39.46%	60.54%
Crafts	1366	2.36%	71.82%	28.18%
Dance	627	1.08%	33.49%	66.51%
Design	4434	7.65%	58.68%	41.32%
Fashion	3454	5.96%	70.64%	29.36%
Film & Video	9818	16.95%	57.83%	42.17%
Food	3871	6.68%	72.18%	27.82%
Games	5113	8.83%	55.21%	44.79%
Journalism	702	1.21%	77.21%	22.79%
Music	8006	13.82%	47.04%	52.96%
Photography	1672	2.89%	66.21%	33.79%
Publishing	6191	10.69%	65.19%	34.81%
Technology	4673	8.07%	74.94%	25.06%
Theater	1796	3.10%	36.69%	63.31%
Grand Total	**57931**	**100.00%**	**59.15%**	**40.85%**

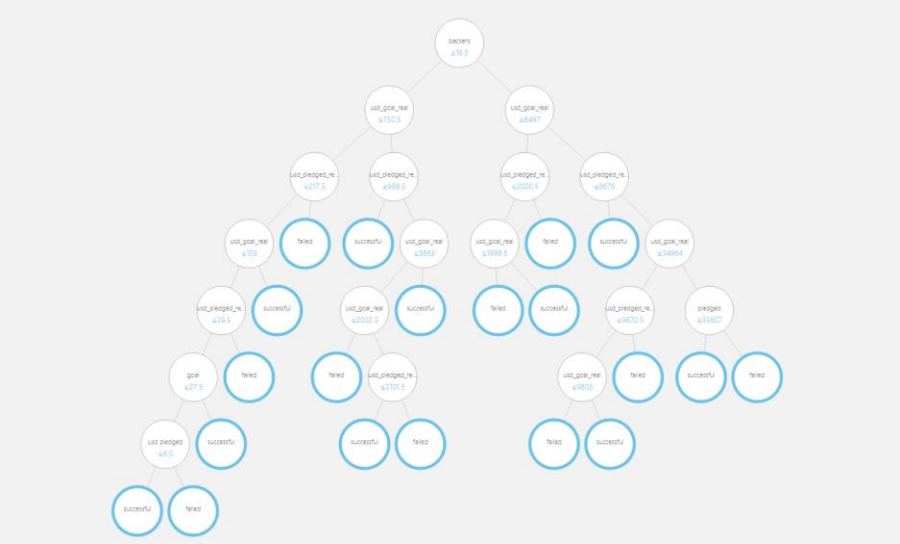
- Pivot table showing the ***Average of state_dummy_fail*** and ***Average of state_dummy_successful*** in Percentage Format.
- Each Percentage describes the fail or success state of the category.
- We can see in the table that comics and theater have high success rate but can't predict the outcome because their count is lower compared to other categories like Film & Video (9818), Music (8006), etc.

Correlation Matrix

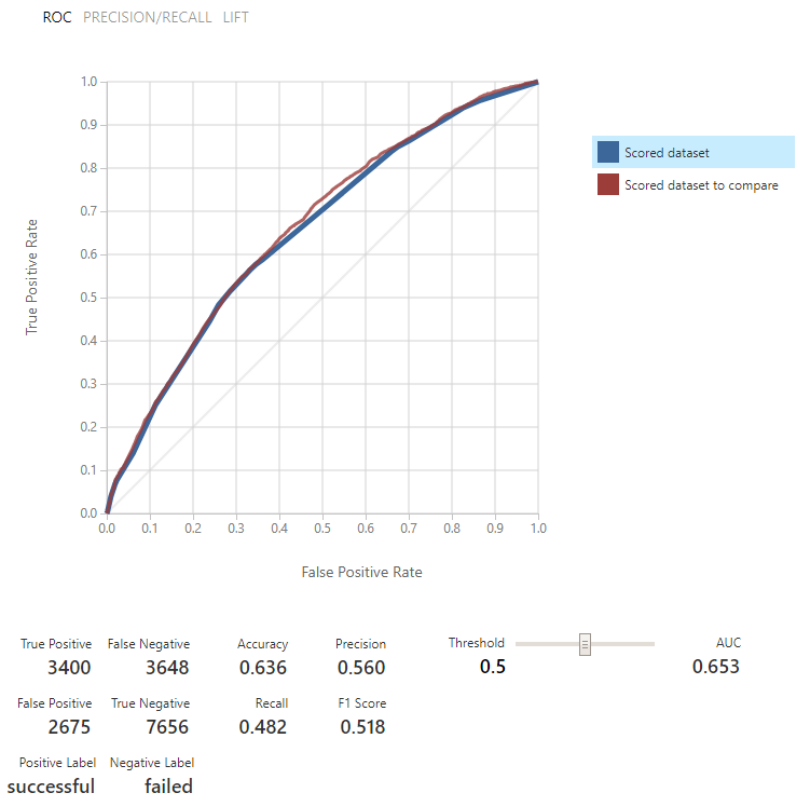
	goal	*pledged*	*backers*	*usd pledged*	*usd_pledged_real*	*usd_goal_real*	*state_dummy_successful*
goal	1						
pledged	0.147059	1					
backers	0.094403	0.688522	1				
usd pledged	0.126598	0.791285	0.675293	1			
usd_pledged_real	0.124936	0.93416	0.751801	0.857111237	1		
usd_goal_real	0.932941	0.120612	0.103065	0.136248831	0.13230851	1	
state_dummy_successful	-0.15379	0.136657	0.149428	0.129212392	0.137348347	-0.155716882	1

1. Correlation Matrix is not a good measure to calculate the outcome because it does not provide us high values for ***state_dummy_successful*** in the given dataset. We move on to the next measure.

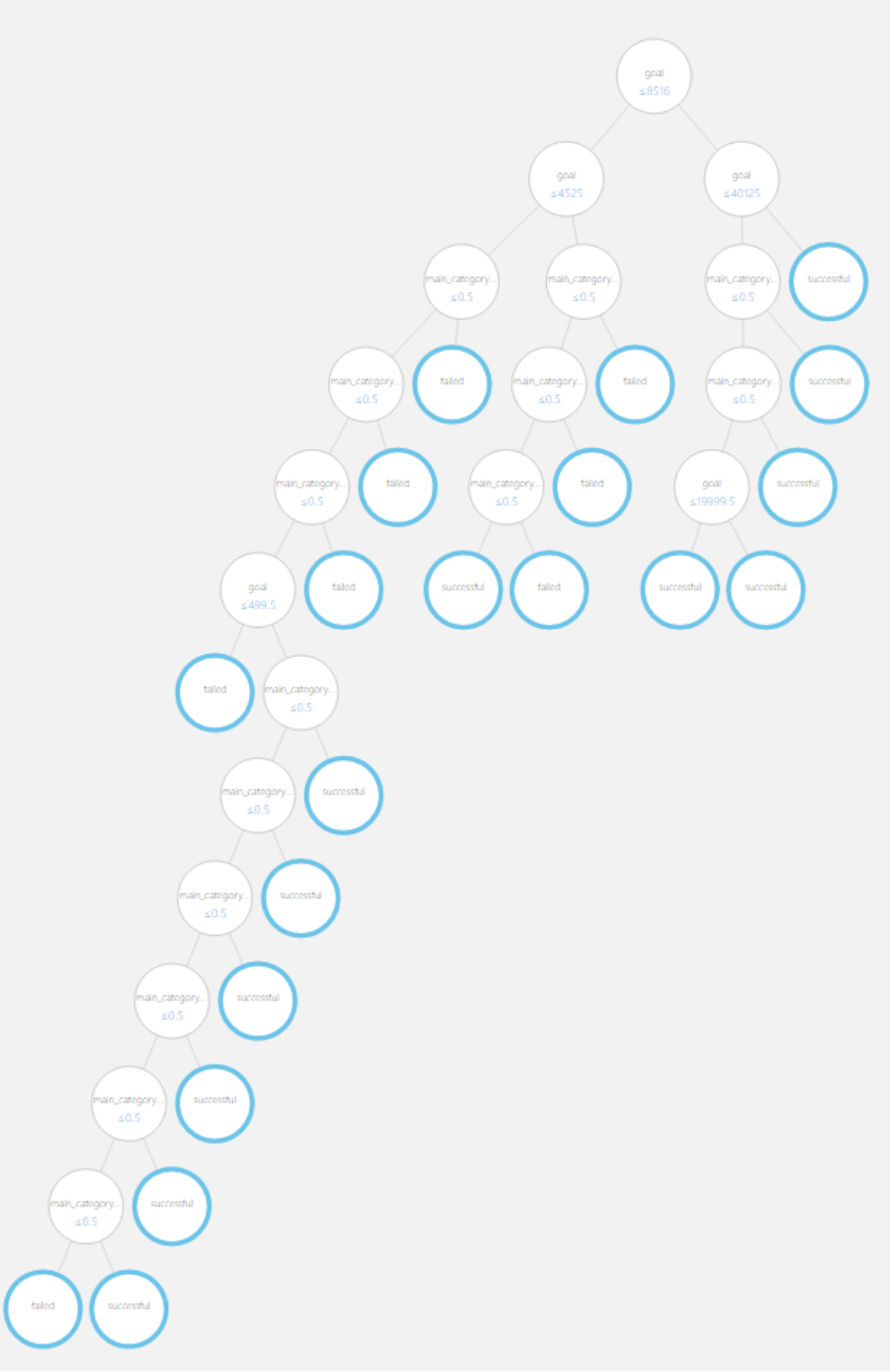
AZURE – Classification model and Regression



- Variables selected for Boosted Decision Tree – main_category, goal, pledged, backers, usd pledged, used_pledged_real, usd_goal_real.
- Trained variable – state.
- Predicting on all the variables in not a good idea because we want to know the success rate before the Kickstarter project has initiated development.



- Variables selected for Logistic Regression – main_category, goal.
- Trained variable - state
- Based on using two variables the 'Accuracy' and 'AUC' seems to be pretty good with this large data set. We will consider the selected variables as good predictors.



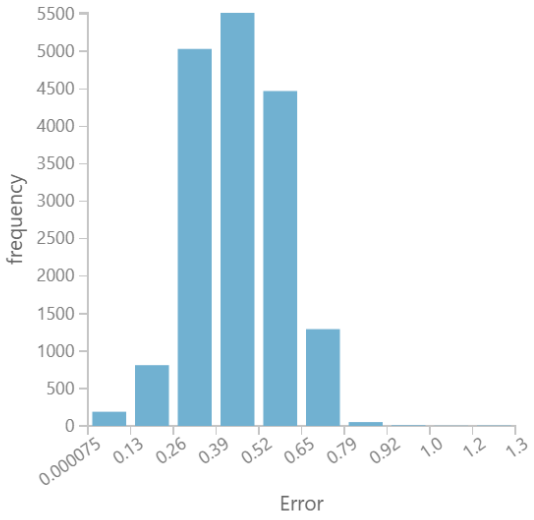
- Variables selected for Boosted Decision Tree – main_category, goal.
- Trained variable – state
- main_category and goal are predictors which can be used before starting the Kickstarter project because the other variables come into play when the project has started development. We would never know their true values before hand. Goal is pre-set by the company to make sure the product has enough funding and they don't suffer loss which leads to cancelation or suspension of the idea.
- If the goal <= 8516 & goal >=4525 & main_catogery is music, result is success.
- If the goal <= 8516 & goal >=4525 & main_catogery is theater, result is success.

Linear Regression

Metrics

Mean Absolute Error	0.454678
Root Mean Squared Error	0.474626
Relative Absolute Error	0.943008
Relative Squared Error	0.934426
Coefficient of Determination	0.065574

Error Histogram

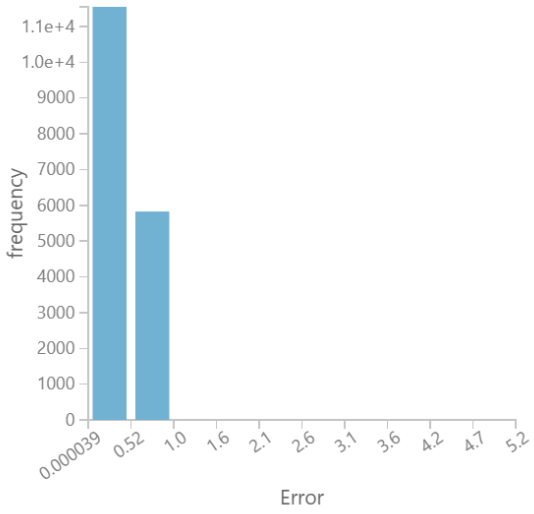


- Variables selected for Linear Regression – main_category, goal.
- Trained variable – state
- Even though the RMSE is 0.474626 is not high but coefficient of 0.065574 is not high either.

Metrics

Mean Absolute Error	0.444008
Root Mean Squared Error	0.465818
Relative Absolute Error	0.920879
Relative Squared Error	0.900064
Coefficient of Determination	0.099936

Error Histogram



- Variables selected for Linear Regression – main_category, goal, pledged, backers, usd pledged, used_pledged_real, usd_goal_real.
- Trained variable – state
- RMSE is 0.465818, coefficient is 0.099936. RMSE is slightly lower and coefficient is much higher compared to the other chart in the figure above.
- This model is better.

Overall Conclusions-

1. Based on the data, choosing the category wisely increases the chances of the success rate of the project. In some cases, like journalism, the trend shows high success for categories which are not very popular and attracts niche customers.
2. Goal is a very important factor. The company needs to set a reasonable goal. This can be determined setting up for a goal with similar categories in the data set.
3. The correlation between the goal and the success can only be determined after the backers have invested in the project but it doesn't guarantee success.

How the results can be used to solve business problem

Kickstarter projects are entirely driven by crowdfunding where the interest of general public and their money is what sends the projects into production. The results can help the forecast of new projects and funders by category for the upcoming year, beneficial to help individuals and starts who wish to launch their idea. It can also provide insight to certain types of media which are more prone to success on the platform.