**Project 1**

Name: Muthu Rishikeshwaran Eswaran

**Executive Summary:**

The main objective of the analysis is to create insights from the available data that can benefit all the parties (i.e. backers, creators and for the platform). Several interaction variables are created in tableau and used for analysis and visualization. The key insights from after analyzing the dataset are:

* The relationship between backers and the total amount pledged is linear and we can help backers (check the information from the tableau visualization) by guiding them to invest in categories which reach has high success rate and performance. (Beneficial for backers)
* Backers are the key to the platform and the key variable in this data. It affects all the key variables positively and from the models and visualizations we can verify that.
* By visualizing the top 3 performing categories grouped by countries we can help backers identify the top categories to invest and reduce the risk.
* By taking the main category, backers and goal we can find the describe the performance of each category and compare the success rate of each category, the amount pledged and the goal.
* By finding out the least performing columns we can show what percentage of risk investing in each category has and what percentage of the goal has been reached.

**Introduction:**

Kickstarter is a website and online fundraising platform on which you can raise funds for a project or new business venture. To raise funds, you describe your project, set a goal for the amount you will raise, and ask “backers” to invest in your project. If you reach your goal, then you receive all the funds received. Otherwise, your project will not be funded, and all the money is returned back to the backers.

The platform is unique and provides various beneficial factors for everyone. I think the most interesting part about the platform is the availability and the ability that anyone can become a Backer or anyone with an idea can become a Creator. The openness and the available data from the platform make it easier for everyone to ask the right questions. Some of the interesting questions that can be analyzed from the data are,

1. Which country has the highest success rate and what is the key reason behind it?
2. Does the platform have enough backers to back the creators and fund the project? What happens when there aren’t enough backers for a project?
3. What is the risk of failure while investing in a category and how to avoid that?
4. Which categories performed consistently throughout the time period of the dataset and what is the reason behind it?
5. What variables in the dataset has good correlation and what models to apply on the dataset?
6. Which categories are performing poorly? What can you say about the subcategories?
7. How do projects that are successful differ from projects that fail? How does this vary with the project category?

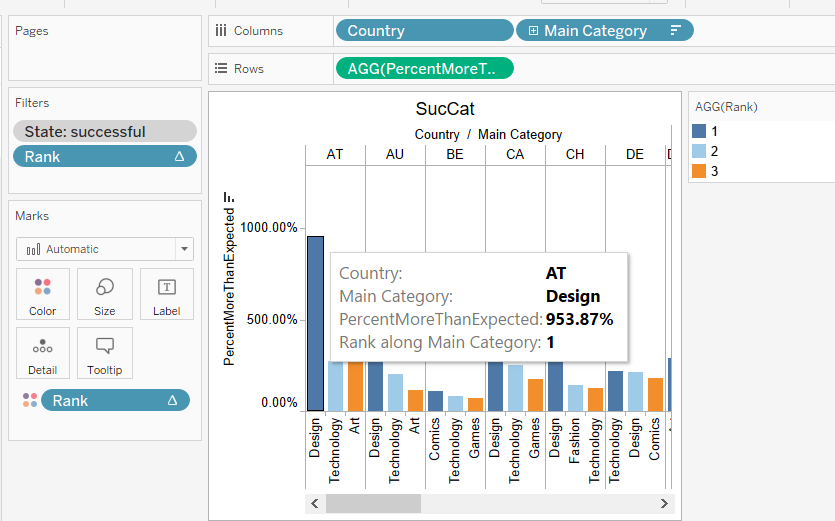
**Insights:**

Insight 1:

The top 3 performing categories in each country is visualized using tableau and from the graph we can identify what category is performing well in each country individually. Also, for these categories the success rate (goal raised) is more than the requirement raised by the creator in all cases for that particular country.

The percentage of fund raised more than the requirement is also calculated and the categories are ranked orderly.

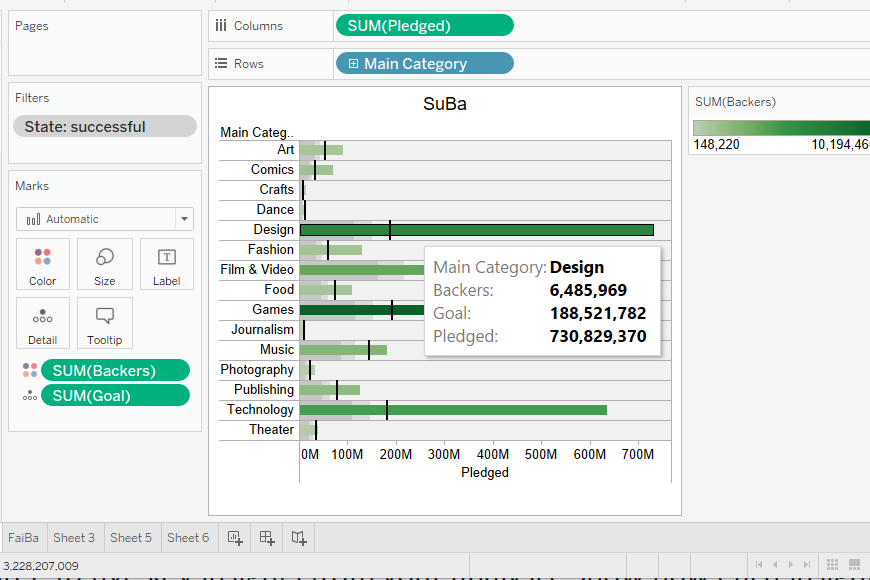
Example: This can help the plat to identify the key contributions for the success of the category.



Insight 2:

The performance of the main categories and the relationship between the backers and the pledged amount is proven here. We can see that as the number of backers increases the pledged amount is increased. Since the platform depends on many backers and not few, the number of backers for a particular project is the most important factor.

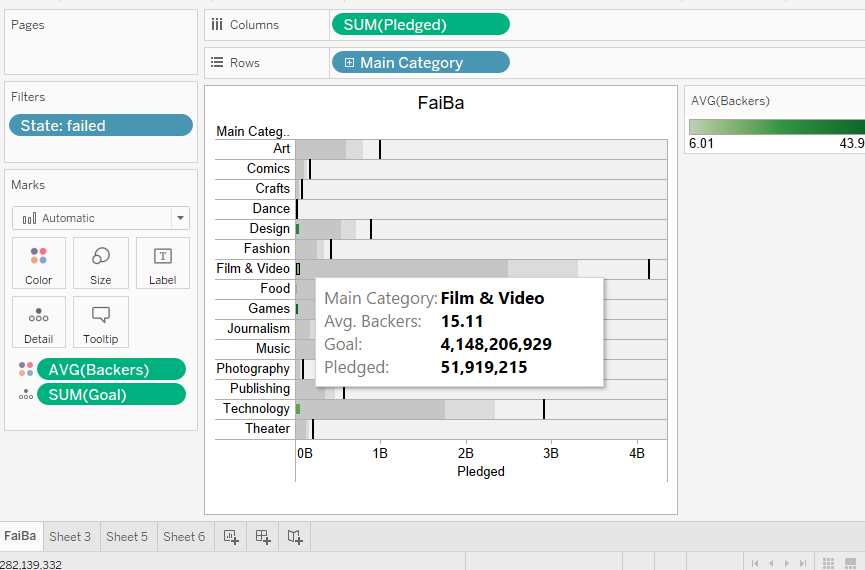
We can also see that design and the technology categories has raised the highest pledged amount and this will give investors a clear idea of where to focus on.



Insight 3:

From the above analysis we concluded that backers are the key part in the platform and higher backers equals higher success rate. Now the projects with low backers are visualized grouping the categories and the pledged amount to check if failure is also mainly dependent on backers.

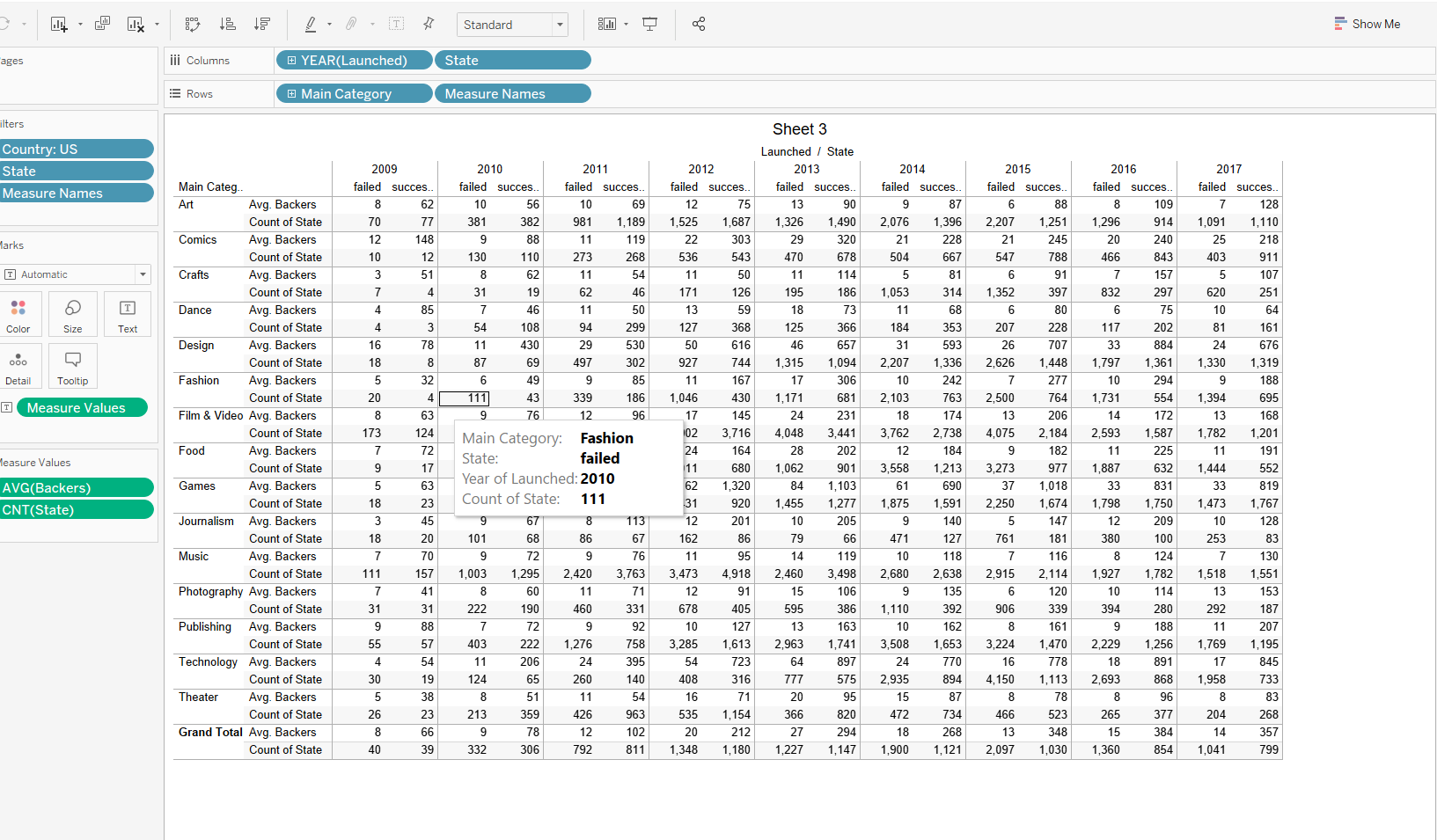
We can verify from the graph that lower backers equal lower success rate. The visualization also displays the actual goal and the amount pledged.

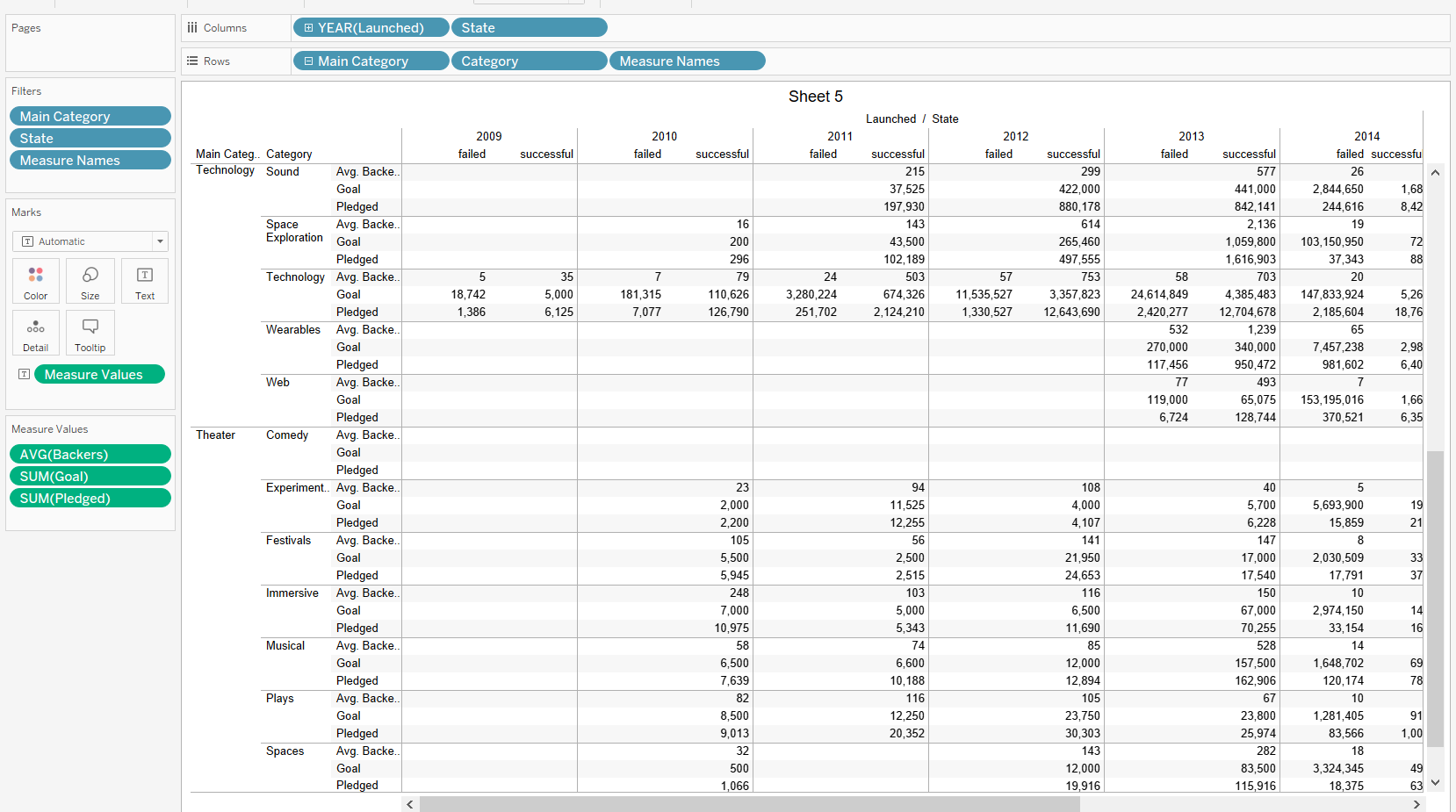


Insight 4

We can identify that US is the biggest user of the platform and decided to infer the top main categories that were successful in US (Technology and Theatre – from Insight 1) and analyzed the number of success and failure counts for each main category with number of backers. Also, described the success and failure status of the sub categories for the years 2009-2017.

A important finding in this analysis is that the sub categories Flight, Gadgets, Sound, Makerspaces, Space Exploration, Wearables, Web (From main category – Technology in US) has a 100% success rate from the years 2009 – 2012 and the sub categories of Theatre in US ever variable except theatre (sub category) has a 100% success rate from the years 2009-2013.





Insight 5:

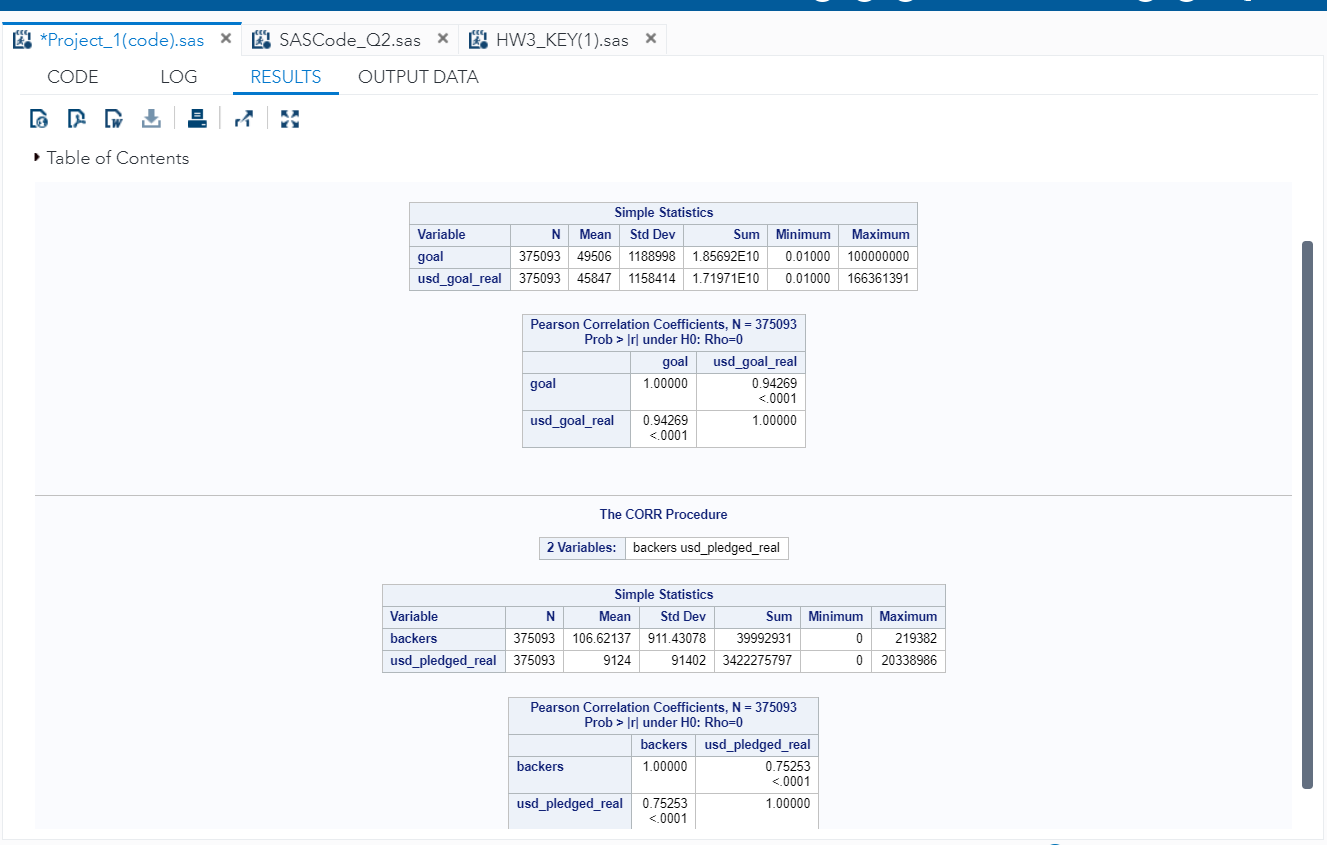
Hypothesis testing is done to check the following condition.

Hypothesis testing:

H0 – Real amount raised is not at least 80% of the required goal.

Ha – Real amount raised is at least 80% of the required goal.

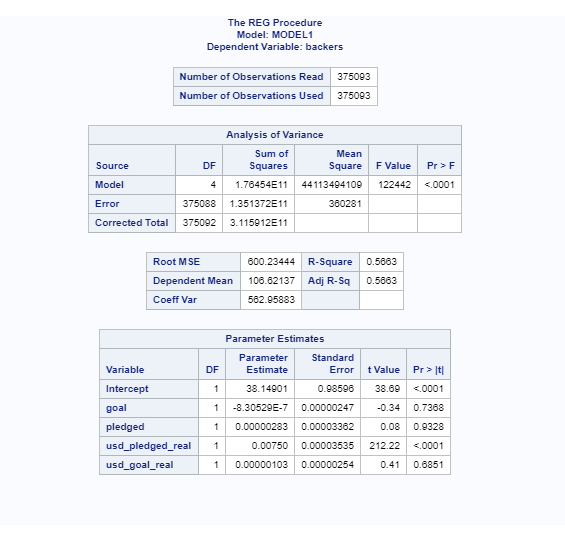
In creator perspective while selecting a goal if the amount raised is at least 80% of the requirement set by the creator then the performance of the platform has an 80%+ success rate. It can attract more creators and help in advertising the platform.



Regression**:**

By focusing on improving backers in the categories where goal is less, we can increase the success rate.

Regression model is done using the backers as target column. Here the linear relationship between backers and usd\_pledged\_real is once again verified.



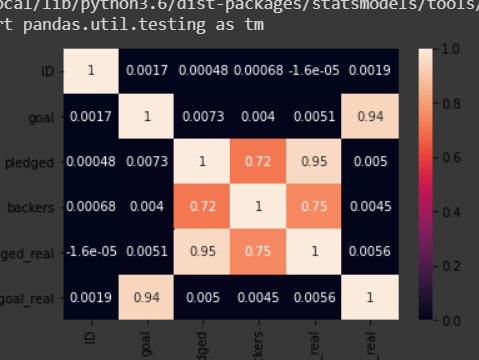
**Limitations and additional suggestions:**

* By one hot encoding the categorical variables we can identify the relation between main categories and sub categories
* By running a time series models we can predict the success rate of the upcoming projects.

**Appendix:**

The given data set is checked for missing variables and the data as provided is clean. The categorical columns are dropped and date time column is separated.

correlation heatmap is plotted for the variables and the most correlated variables are saved in a new data frame and exported as csv.



Platform used: GoogleColab

Language used: Python

Code:

import pandas as pd

import numpy as np

df[['date\_launched','time\_launched']] = df['launched'].str.split(' ',expand = True)

df.isna()

df.drop(['launched','time\_launched'],axis =1)

df1 = df.drop(['ID','name','category','main\_category','currency','deadline','state','country','date\_launched'],axis=1)

import seaborn as sn

import matplotlib.pyplot as plt

corrMatrix = df.corr()

sn.heatmap(corrMatrix, annot=True)

plt.show()

df1.to\_csv("kick.csv", index=False)

The file is then uploaded in SAS and hypothesis testing is done for the condition using the variables selected in Googlecolab. Then regression is done for the data using the following code:

data kick;

infile '/folders/myfolders/project\_1/kick.csv' dlm=',' firstobs=2 dsd;

informat goal best12.;

informat launched ddmmyyyyw.;

informat pledged best12.;

informat backers best12.;

informat usd\_pledged\_real best12.;

informat usd\_goal\_real best12.;

informat time\_launched ;

input goal launched pledged backers usd\_pledged\_real usd\_goal\_real time\_launched;

data kick;

set kick;

drop launched time\_launched;

/\* Hypo testing \*/

proc sort data = kick;;

by goal;

run;

proc corr data=kick;

var goal usd\_goal\_real ;

run;

proc corr data=kick;

var backers usd\_pledged\_real ;

run;

/\* Regression \*/

proc reg data = kick;

model backers = goal pledged usd\_pledged\_real usd\_goal\_real ;

run;

Exploratory analysis and visualization are carried out in tableau for all numerical and categorical data. Very complex questions and key insights are obtained from the tableau analysis. Several interaction variables are created and visualized.