

A Closer Look @Projects on KICKSTARTER

Dream Maker or Promise Breaker



Asha Justin & Priyanka Sabnani

TABLE OF CONTENTS

Problem statement.....	2
Project Description.....	2
Business uses from the project.....	3
Data set.....	4
Model 1.....	7
Model 2.....	18
Analysis and Graphs.....	31
Challenges.....	36
Conclusion.....	36
Lessons Learned.....	37
References.....	37

Problem statement

In recent years, the range of funding options for projects created by individuals and small companies has expanded considerably. Kickstarter, American public benefit corporation based in New York mission is to "Help bring creative projects to life". As of December 2019, Kickstarter has received more than \$4.6 billion in pledges from 17.2 million backers to fund 445,000 projects.

Kickstarter has grown massively since its launch. Kickstarter management team now wants to create a database which can track details about their creators, backers and projects. They would like to track who are people creating projects with them, who are the people backing these projects and the performance of their company.

Project Description

A huge variety of factors contribute to the success or failure of a project. Some of these can be quantified or categorized, which allows for the construction of a model to attempt to predict whether a project will succeed or not.

The aim of this project is to construct such a model and to analyze Kickstarter project data more generally, in order to help potential project creators assess whether Kickstarter is a good funding option for them, and what their chances of success are. This project will also help the management at Kickstarter to analyze the health of the companies by knowing about the creators, backers and their trend in investing money in types of projects created at Kickstarter.

BUSINESS USES FROM THE PROJECT

Kickstarter is main sources of initial capital for small businesses and start-up companies that are looking to launch their first products. This project will help them understand trending projects in market with their success rates.

This project will help the management team predict the success or failure of the project before its released. It will also help them understand about the creators and backers assigned with various types of projects. This project will help the management team to make critical decisions which will support in the growth of their organization.

What can the data tell us?

- Most successful projects by Country or Category
- Average Goal Amount for a successful project by Category Creators with most successful projects
- Having Facebook connection /website effect success of a project
- Which categories have projects with maximum profit Who are the people creating projects in different categories
- What are the reasons for success / failure of the projects?

DATA SET

Kickstarter campaigns make ideas into reality. It is where creators share new visions for creative work with the communities that will come together to fund them.

Kickstarter data set contains:

- Data set contains details for 18000+ unique projects
- Project details by the countries they were developed
- Details about creators and backers of the project
- Details about project with details of Amount Funded Vs the Goal Amount
- Project details based on Category & Subcategory
- Project details based on their status and profit / loss
- Details about Creators and the rewards they have received from individual projects
- Details about the start and end date of the projects

DATA IMPORT

How we arrived at our final database:

Raw data for the report was downloaded from Kaggle. We analyzed different datasets with under Kickstarter to get the best dataset to analyze. We analyzed the raw dataset and removed columns which would not help us analyze any point. We created unique ids for City and Creator to ensure there are no duplications and the data can be projected as given. We analyzed the to understand its form and then transformed it to 2NF and finally to 3NF.

Data Warehouse Design - Schema

Dimension Table

Fact Table

DimensionTable

Project

ProjectID (PK)

Name

URL

Project State

Description

Category

Location

LocationID (PK)

City

State

Country

Revenue

ProjectID (PK) (FK)

LocationID (PK) (FK)

CreatorID(PK) (FK)

DateID (PK) (FK)

Goal Amount

Pledged Amount

Duration

Facebook Shares

Creator - # Projects Created

Creator - # Projects Backed

Creator

CreatorID (PK)

Name

Facebook Connected

Facebook Friends

Creator Website

Date

DateID (PK)

StartDate

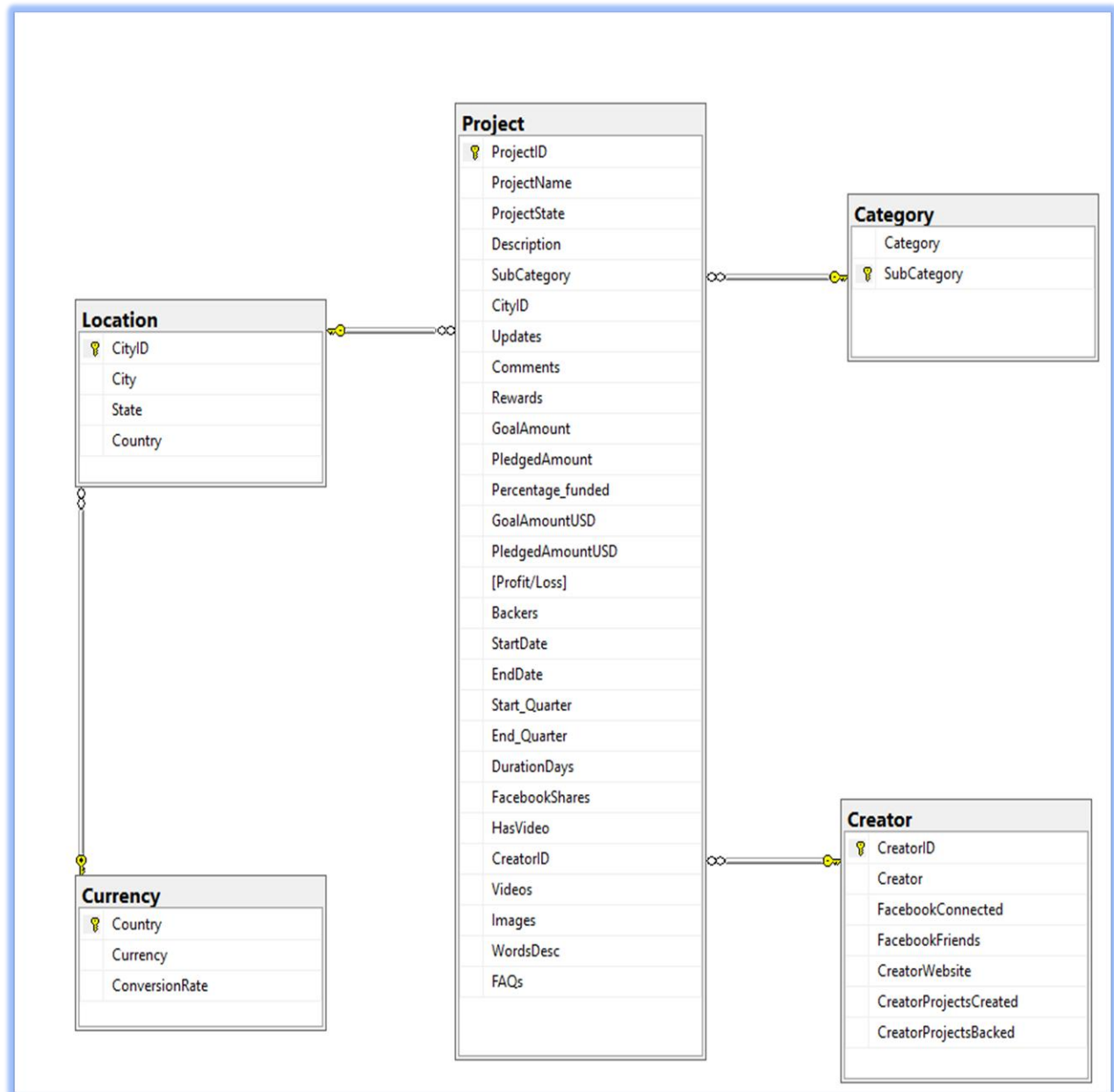
EndDate

Month

Year

Quarter

DATAWARE HOUSE MODEL – SQL



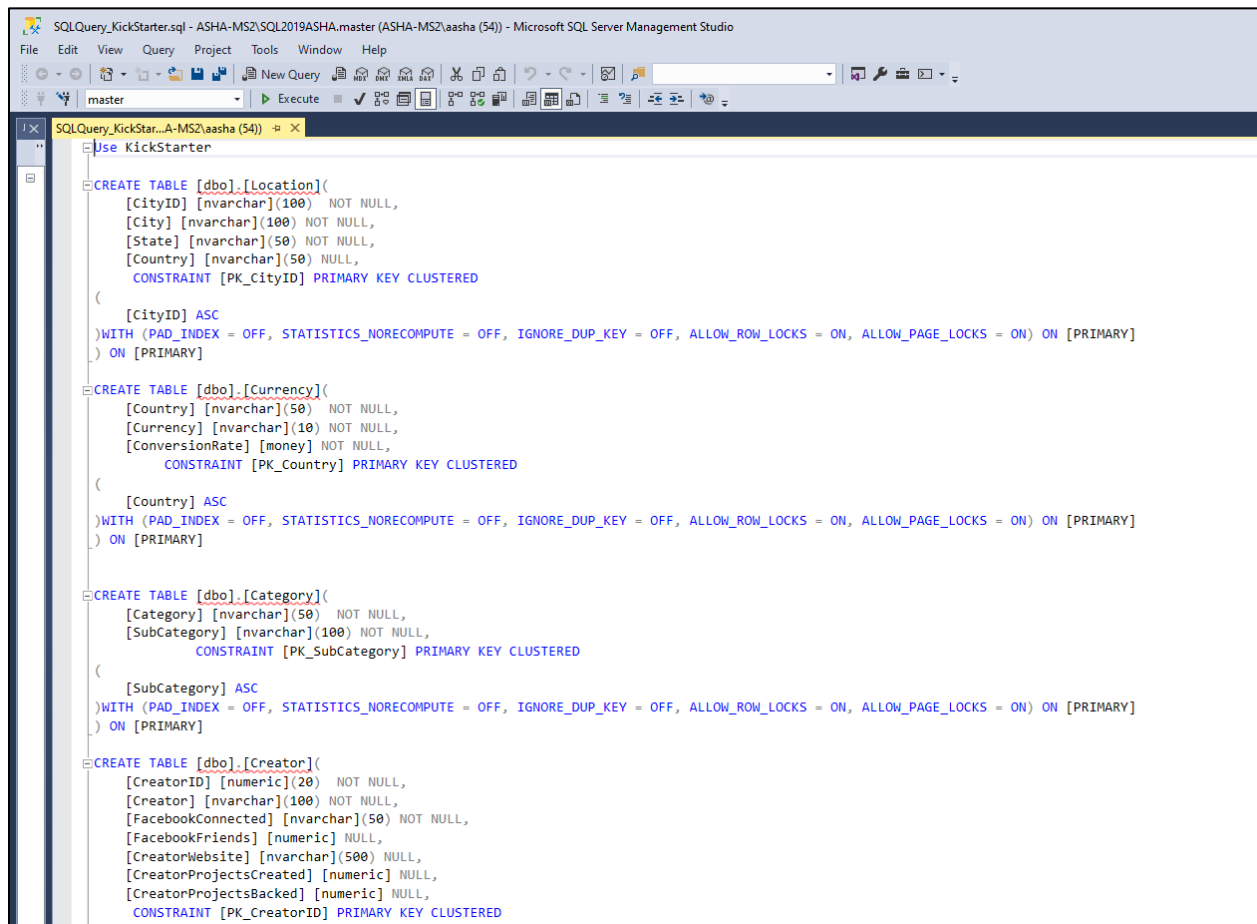
MODEL

MODEL 1 – Visual Studio

Model was built in Visual Studio by creating a data warehouse, and extracting data into SQL using Integration Services Project. After the data is loaded into SQL, we created cubes and reports to analyze the data. This analysis will help the management team understand which countries or categories invest most in Kickstarter projects. This will also help them understand various reasons for success and failure of the projects.

Extraction Transformation Load

Created tables (Destination) in SQL where the data extracted from the excel file (data source) can be loaded.



```
SQLQuery_KickStarter.sql - ASHA-MS2\SQL2019ASHA.master (ASHA-MS2\asha (54)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

master Execute

SQLQuery_KickStar...A-MS2\asha (54)

Use KickStarter

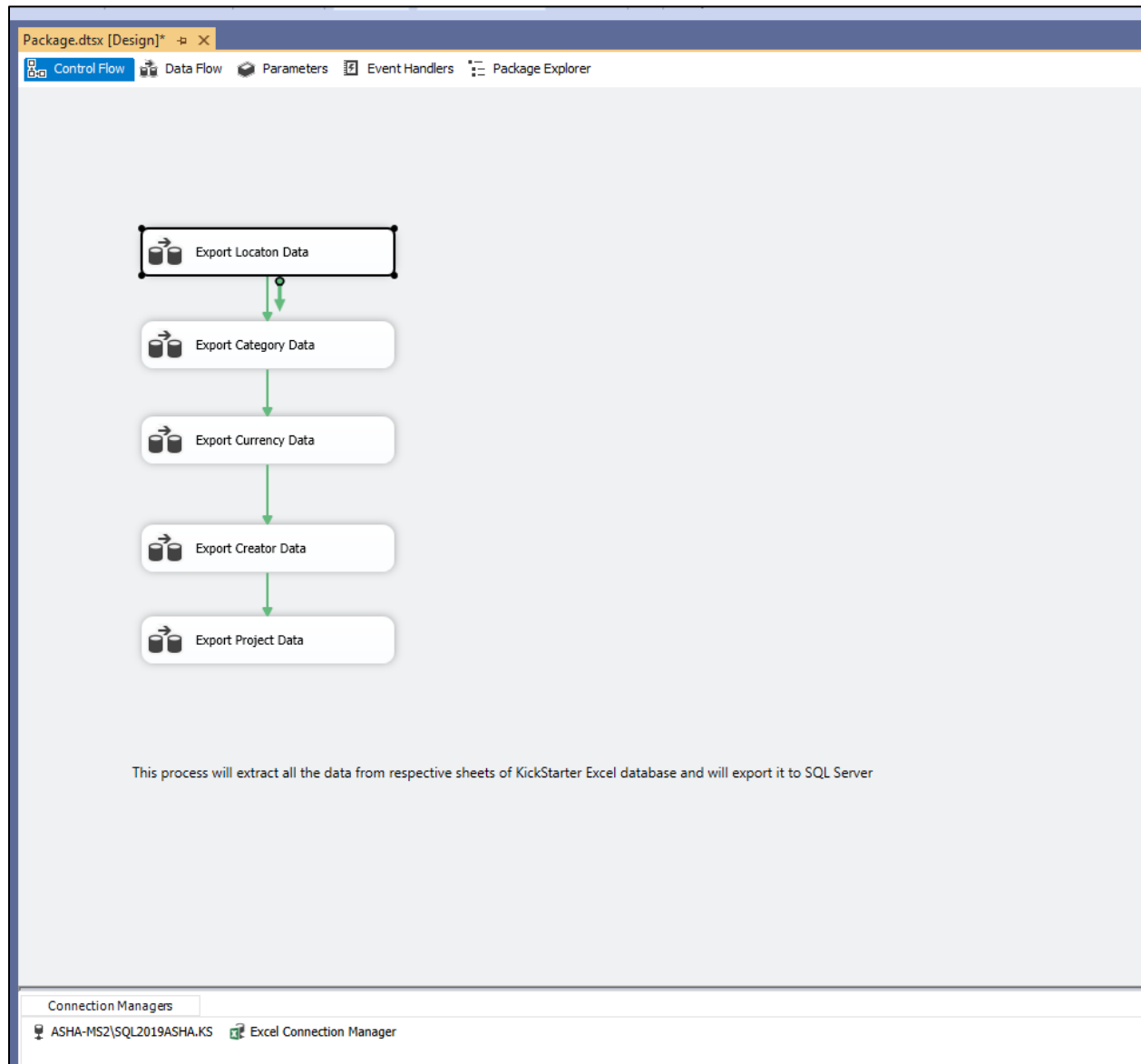
CREATE TABLE [dbo].[Location](
    [CityID] [nvarchar](100) NOT NULL,
    [City] [nvarchar](100) NOT NULL,
    [State] [nvarchar](50) NOT NULL,
    [Country] [nvarchar](50) NULL,
    CONSTRAINT [PK_CityID] PRIMARY KEY CLUSTERED
(
    [CityID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[Currency](
    [Country] [nvarchar](50) NOT NULL,
    [Currency] [nvarchar](10) NOT NULL,
    [ConversionRate] [money] NOT NULL,
    CONSTRAINT [PK_Country] PRIMARY KEY CLUSTERED
(
    [Country] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

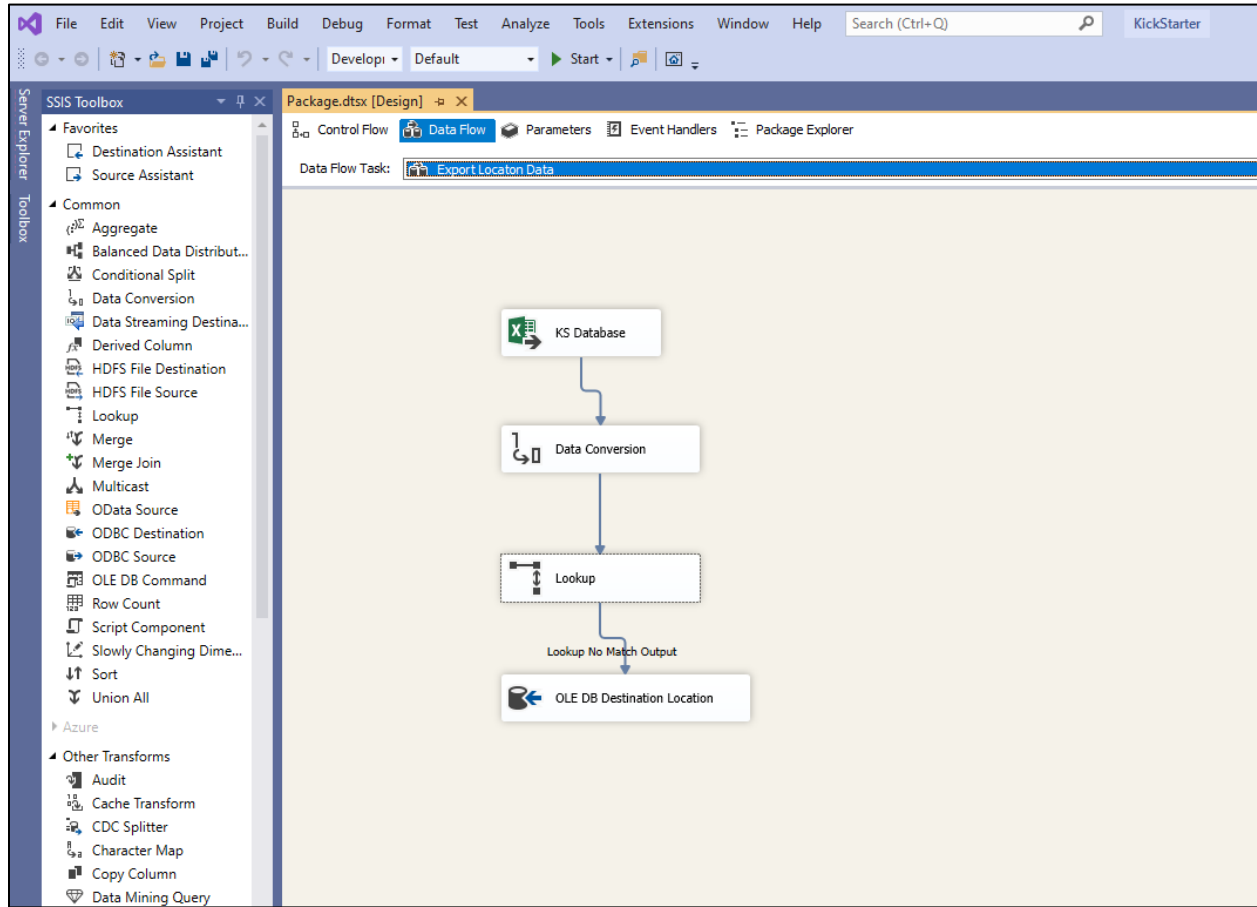
CREATE TABLE [dbo].[Category](
    [Category] [nvarchar](50) NOT NULL,
    [SubCategory] [nvarchar](100) NOT NULL,
    CONSTRAINT [PK_SubCategory] PRIMARY KEY CLUSTERED
(
    [SubCategory] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[Creator](
    [CreatorID] [numeric](20) NOT NULL,
    [Creator] [nvarchar](100) NOT NULL,
    [FacebookConnected] [nvarchar](50) NOT NULL,
    [FacebookFriends] [numeric] NULL,
    [CreatorWebsite] [nvarchar](500) NULL,
    [CreatorProjectsCreated] [numeric] NULL,
    [CreatorProjectsBacked] [numeric] NULL,
    CONSTRAINT [PK_CreatorID] PRIMARY KEY CLUSTERED
```


Created ETL package using SSIS to move data from Excel data source to SQL tables using Integration services project. Created control flow for all the tables and the flow they will extract data into the tables. Each of the tables had separate data flow associated. Data flow had extraction, data conversion, lookup for new data (if available) and loading data to SQL tables. Data flow was managed by adding required connections to the source.

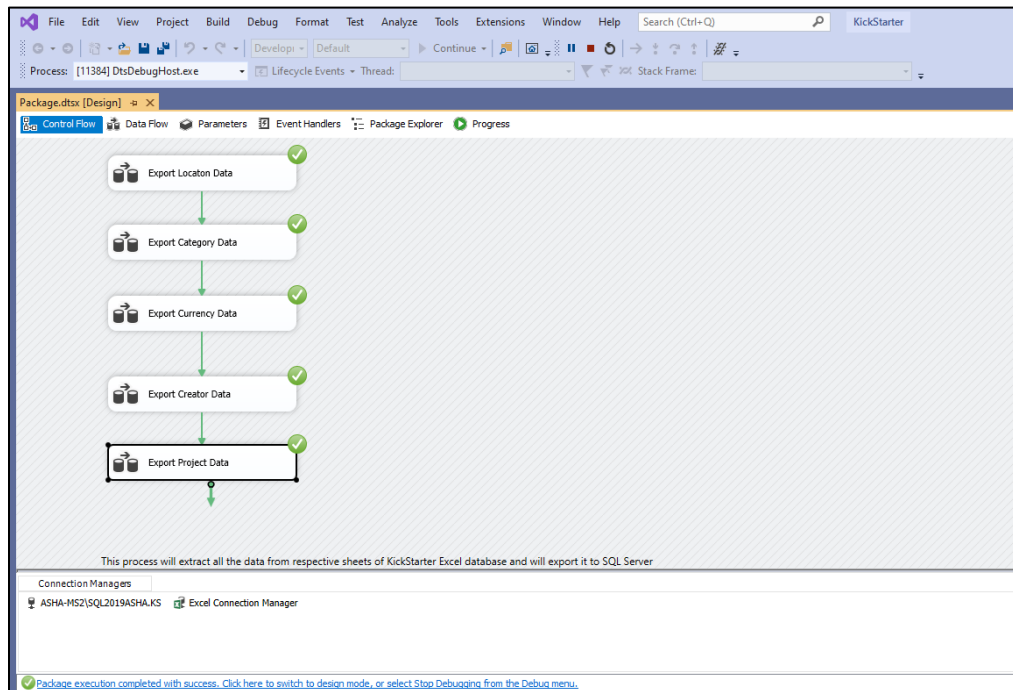


Data Flow – Location Table



Similarly created different data flow tasks for all the tables in SQL

Execute the all the data flow by clicking on the start button and execute the control flow. Once the control flow is executed data from excel will be uploaded to SQL



SQLQuery1.sql - ASHA-MS2(SQL2019ASHA)KS (ASHA-MS2\asha (64)) - Microsoft SQL Server Enterprise Manager

File Edit View Query Project Tools Window Help

Connect - KS

Object Explorer

ASHA-MS2(SQL2019ASHA) (SQL Server 15.0.2070.41 - ASHA-MS2\asha)

Databases

System Databases

Database Snapshots

AdventureWorksDW2012

AdventureWorksDW2017

Bakery

ContractsMidterm

CornerMed_Justin

GemDataWareHouse

KickStarter

KS

Database Diagrams

Tables

System Tables

FileTables

External Tables

Graph Tables

dbo.Category

dbo.Creator

dbo.Currency

dbo.Location

dbo.Project

Views

External Resources

Synonyms

Programmability

Service Broker

Storage

Security

Movies

Petstore

ReportServer

ReportServerTempDB

Sample

SSIS_Demo

Security

Server Objects

Replication

PolyBase

Always On High Availability

Management

Integration Services Catalogs

SQL Server Agent

XEEvent Profiler

SQLQuery1.sql - ASHA-MS2\asha (64)*

Use KS

select * from Location;

Results Messages

110 %

CityID	City	State	Country
AK1	Willow	Alaska	United States of America
AK10	Wasilla	Alaska	United States of America
AK2	Fairbanks	Alaska	United States of America
AK3	Sitka	Alaska	United States of America
AK4	Anchorage	Alaska	United States of America
AK5	North Pole	Alaska	United States of America
AK6	Sutton	Alaska	United States of America
AK7	Eagle River	Alaska	United States of America
AK8	Juneau	Alaska	United States of America
AK9	Palmer	Alaska	United States of America
AL1	Auburn	Alabama	United States of America
AL10	Tuscaloosa	Alabama	United States of America
AL11	Montgomery	Alabama	United States of America
AL12	Gulf Shores	Alabama	United States of America
AL13	Fairhope	Alabama	United States of America
AL14	Muscle Shoals	Alabama	United States of America
AL15	Cullman	Alabama	United States of America
AL16	Heflin	Alabama	United States of America
AL17	Prichard	Alabama	United States of America
AL18	Dothan	Alabama	United States of America

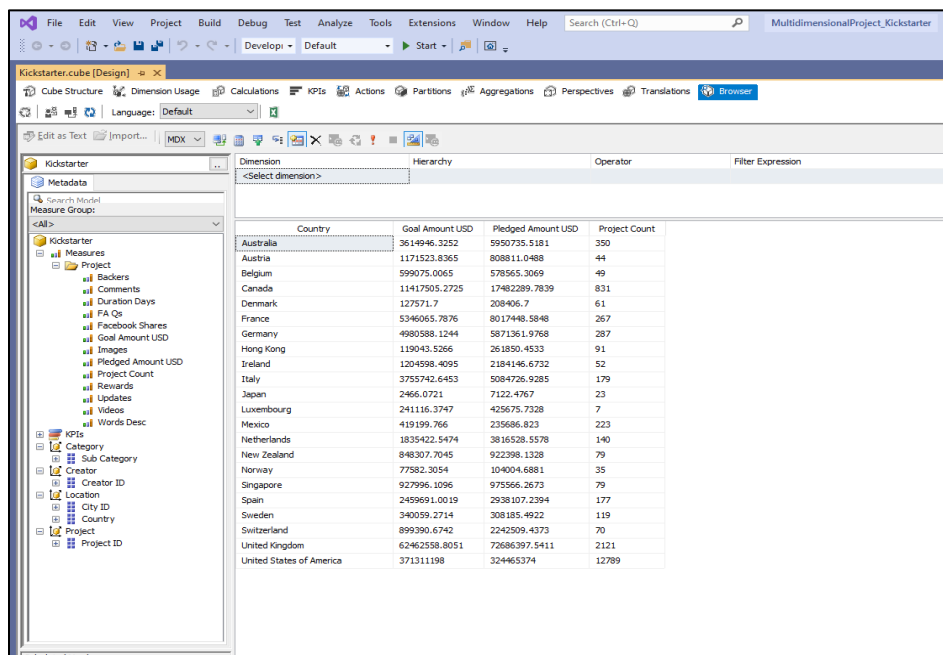
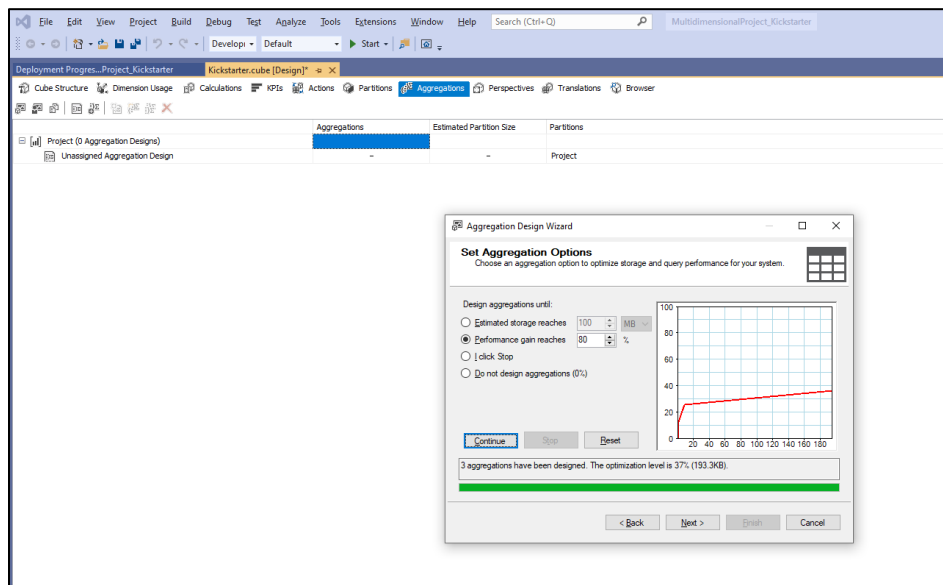
Query executed successfully.

ASHA-MS2(SQL2019ASHA) (15.0.2070.41)

Data Analysis

SSAS-

Created cubes using the data exported to SQL to analyze data. Added data source from SQL, created data source views and then created a cube based on the data source view. Processed the cube to generate the final output. Once the cube is processed click on browser tab and add measures and dimensions as required.



Data Analysis using Cubes – Added new dimensions to cube and re-processed the cube to analyze requirements given. Wrote MDX queries to sort the results from cube. We can add as many dimensions required to get inputs from the data.

The screenshot shows the SQL Server Enterprise Manager interface for the 'Kickstarter.cube [Design]'. The left pane displays the metadata tree, including dimensions like Country, Project, and Project Count. The right pane shows the MDX query and a table of results.

```

SELECT
  NON EMPTY
  {Measures.[Project Count]} ON COLUMNS,
  Order(
    {Location].[Country].[Country].MEMBERS
  ),Measures.[Project Count]
  ,DESC
)
FROM [Kickstarter].[CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS]

```

Country	Project Count
United States of America	12789
United Kingdom	2121
Canada	831
Australia	350
Germany	287
France	267
Mexico	223
Italy	179
Spain	177
Netherlands	140
Sweden	119
Hong Kong	91
New Zealand	79

The screenshot shows the SQL Server Enterprise Manager interface for the 'Kickstarter.cube [Design]'. The left pane displays the metadata tree, including dimensions like Country, Project, and Project Count. The right pane shows the MDX query and a table of results.

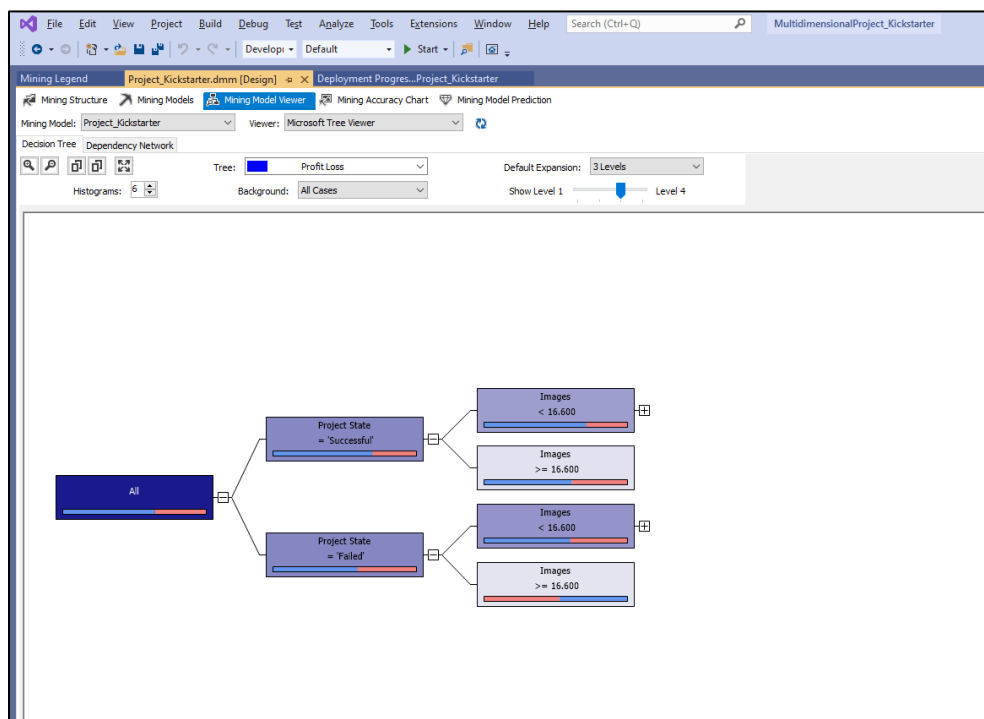
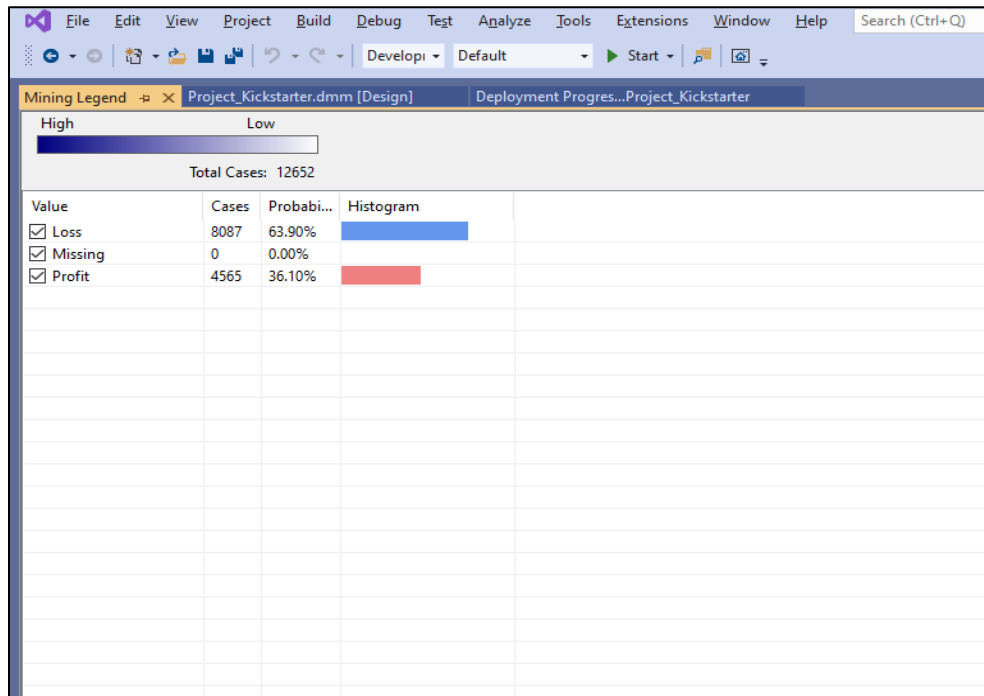
```

SELECT
  NON EMPTY
  {Measures.[Project Count]} ON COLUMNS,
  Order(
    {Location].[Country].[Country].MEMBERS
  ),Measures.[Project Count]
  ,DESC
)
FROM [Kickstarter].[CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS]

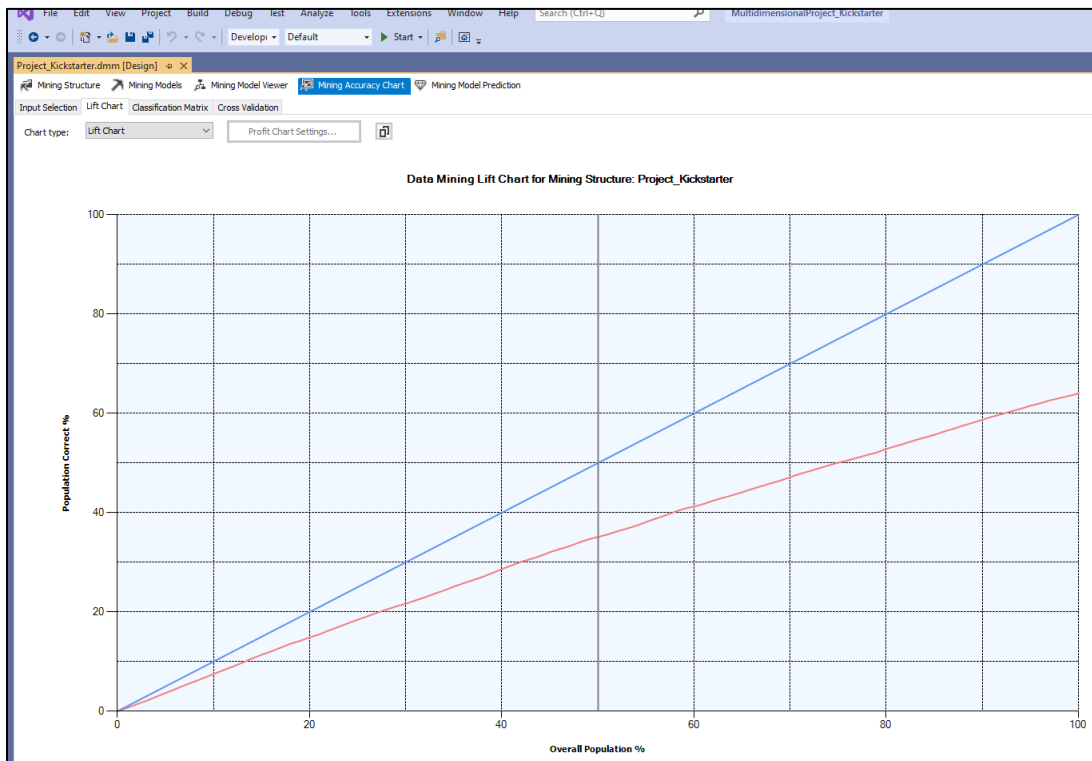
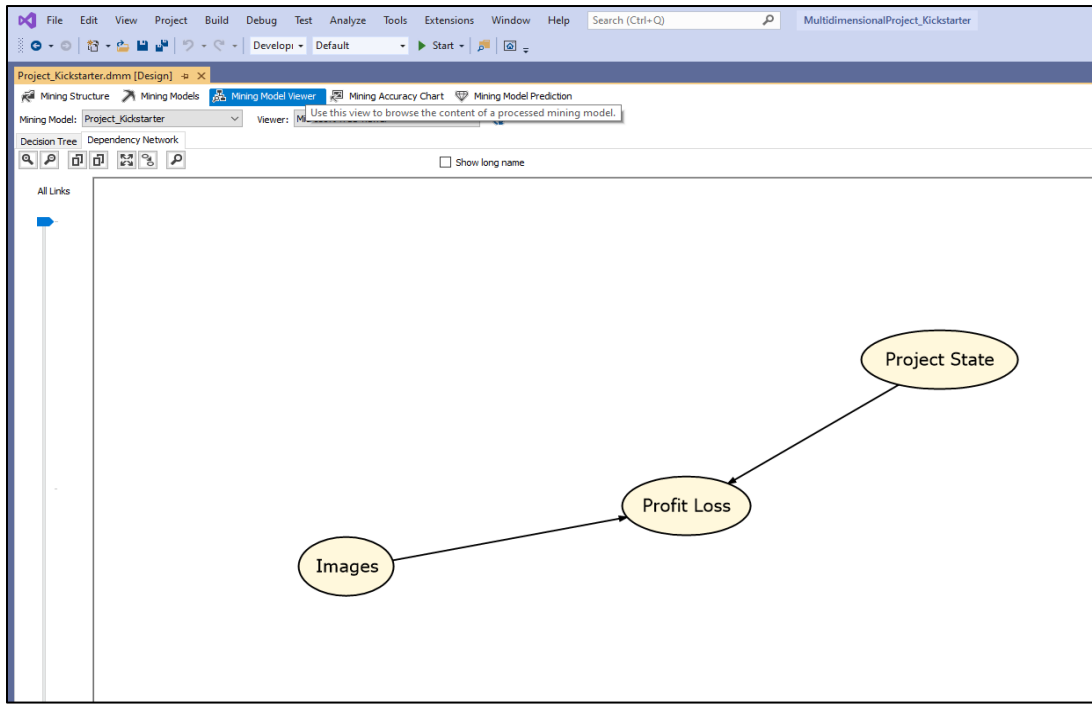
```

Country	Profit Loss	Project Count
Australia	Loss	220
Australia	Profit	130
Austria	Loss	27
Austria	Profit	17
Belgium	Loss	28
Belgium	Profit	21
Canada	Loss	530
Canada	Profit	301
Denmark	Loss	41
Denmark	Profit	20
France	Loss	173
France	Profit	94
Germany	Loss	168
Germany	Profit	119
Hong Kong	Loss	64
Hong Kong	Profit	27
Ireland	Loss	32
Ireland	Profit	20
Italy	Loss	108
Italy	Profit	71
Japan	Loss	18
Japan	Profit	5
Luxembourg	Loss	4
Luxembourg	Profit	3
Mexico	Loss	141
Mexico	Profit	82
Netherlands	Loss	85
Netherlands	Profit	55
New Zealand	Loss	45
New Zealand	Profit	34
Norway	Loss	22
Norway	Profit	13

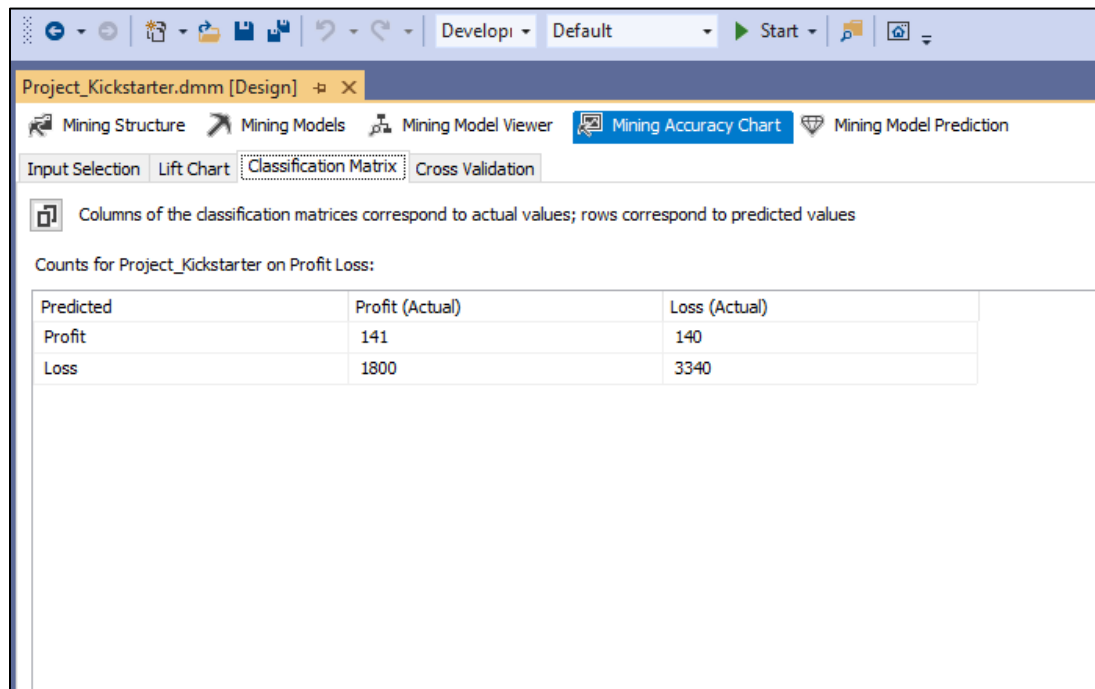
Added Data Mining Wizard to add data prediction to our model. The model can predict the output using input predictors about the success of the model. Decision trees and mining legends are the models that can be used for prediction the success rate.



We can also view the dependent input variables from the dependency network in Mining model viewer. Lift curve will give details about how good the predictors are to predict the output.



Lift Curve will also predict the output based upon the class of the output.

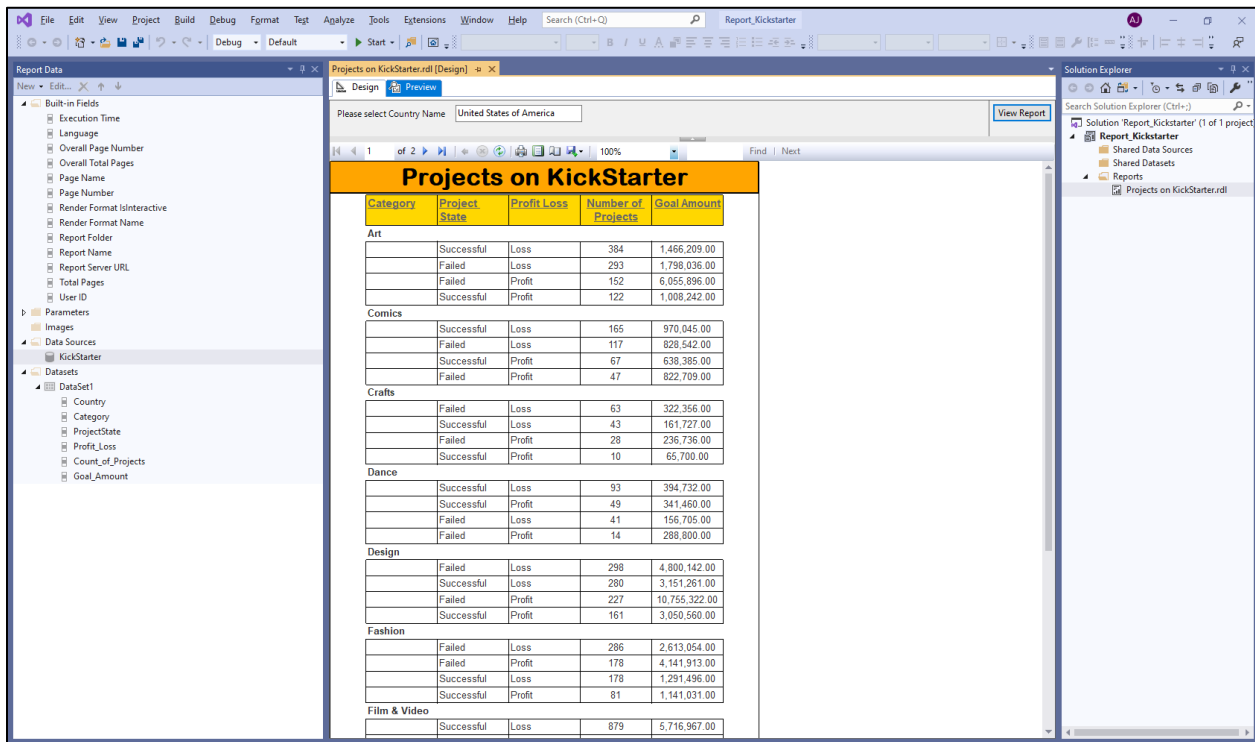
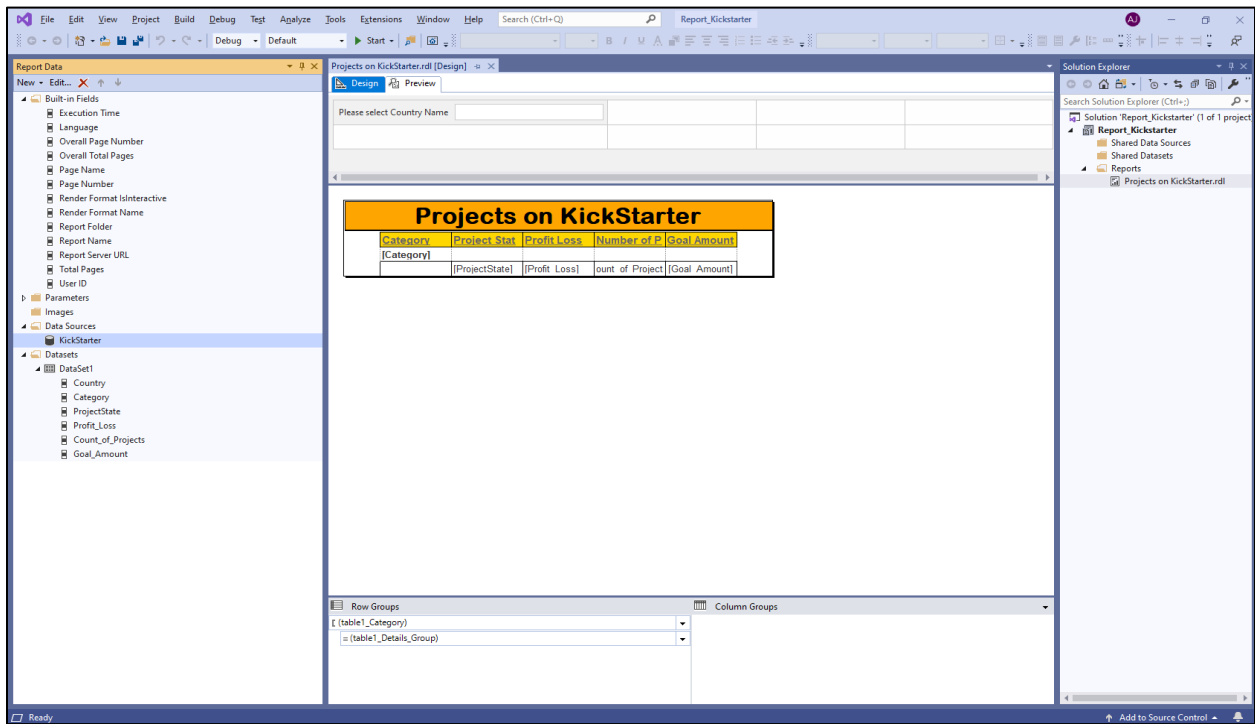


Analysis from Cubes Created

- United States of America has most projects as compared to other countries on Kickstarter
- US also sees a good number in terms of Goal Amount and Pledged amount
- Profit & Loss status by each country
- Images uploaded by Country – US again see a spike in images uploaded
- Categories with Maximum projects – Film Video has the maximum number of projects and Journalism & craft do not see a good number
- Days taken to complete the project by Category – Film Video, Music projects take long to complete but Journalism projects are shortest
- Backers by each project category – Games category sees a high number in Backers (Most Popular among people)

SSRS –

We also created static report using reporting services in Visual Studio which can be uploaded to SharePoint and the team can access data with appropriate access.



Report was then deployed to SharePoint

Course: MIS-6060-12407 / MIS-4 x | Inbox - asjustin@oakland.edu - C x | localhost/ReportServer - / x +

localhost/ReportServer

Apps Oakland Other

localhost/ReportServer - /

Wednesday, February 5, 2020 8:40 PM	<dir> Report Project AdventureWorks
Friday, February 7, 2020 2:06 PM	<dir> Report Project Bakery
Wednesday, February 5, 2020 7:54 PM	<dir> Report Project Sample
Friday, February 7, 2020 12:08 PM	<dir> Report Project SampleJustin
Monday, February 17, 2020 2:05 PM	<dir> Report Project1
Monday, February 17, 2020 10:51 AM	<dir> Report Project1Practice1
Monday, February 17, 2020 11:10 AM	<dir> Report Project1Practice2
Tuesday, April 7, 2020 3:52 PM	<dir> Report Kickstarter
Wednesday, February 19, 2020 7:40 PM	<dir> W20_Justin_Report

Microsoft SQL Server Reporting Services Version 15.0.1102.675

Projects on Kickstarter - Report x +

localhost/ReportServer/Pages/ReportViewer.aspx?%2fReport_Kickstarter%2fProjects+on+Kickstarter&rs:Command=Render

Apps Oakland Other

Please select Country Name: [View Report](#)

1 of 2 ? 100% Find | Next

Projects on Kickstarter

Category	Project State	Profit Loss	Number of Projects	Goal Amount
Art				
	Failed	Loss	21	54,983.16
	Successful	Loss	17	38,437.64
	Successful	Profit	10	21,747.61
	Failed	Profit	7	405,040.23
Comics				
	Successful	Loss	15	48,732.71
	Failed	Loss	5	51,659.61
	Failed	Profit	4	36,486.86
	Successful	Profit	4	7,925.96
Crafts				
	Failed	Loss	5	7,777.84
	Successful	Loss	3	9,501.03
Dance				
	Successful	Loss	3	4,371.20
	Failed	Profit	3	40,605.17
	Successful	Profit	2	8,814.65
	Failed	Loss	2	3,973.82
Design				
	Failed	Loss	28	176,184.55
	Successful	Loss	17	153,100.29
	Successful	Profit	13	185,689.20
	Failed	Profit	9	302,573.58
Fashion				
	Failed	Loss	15	94,179.44
	Successful	Loss	15	56,066.93
	Failed	Profit	12	78,753.81
	Successful	Profit	10	117,119.20
Film & Video				
	Successful	Loss	64	357,627.54
	Failed	Loss	39	379,729.91
	Successful	Profit	27	273,126.88
	Failed	Profit	26	1,641,110.14

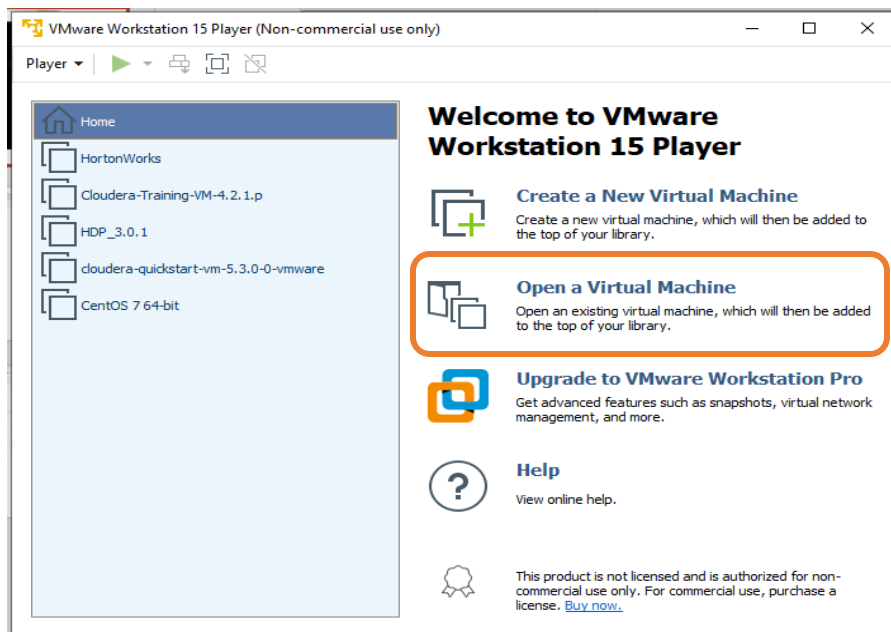
MODEL 2 – Horton Sandbox and MS Power BI

Model is built in Horton Sandbox, by creating data warehouse, and extracting data into Hive by performing map reduce jobs. Data obtained was first exported to GitHub and was then imported Horton Sandbox environment. Processed data was then moved to HDFS and then exported to MS PowerBI to create graphs using ODBC connection.

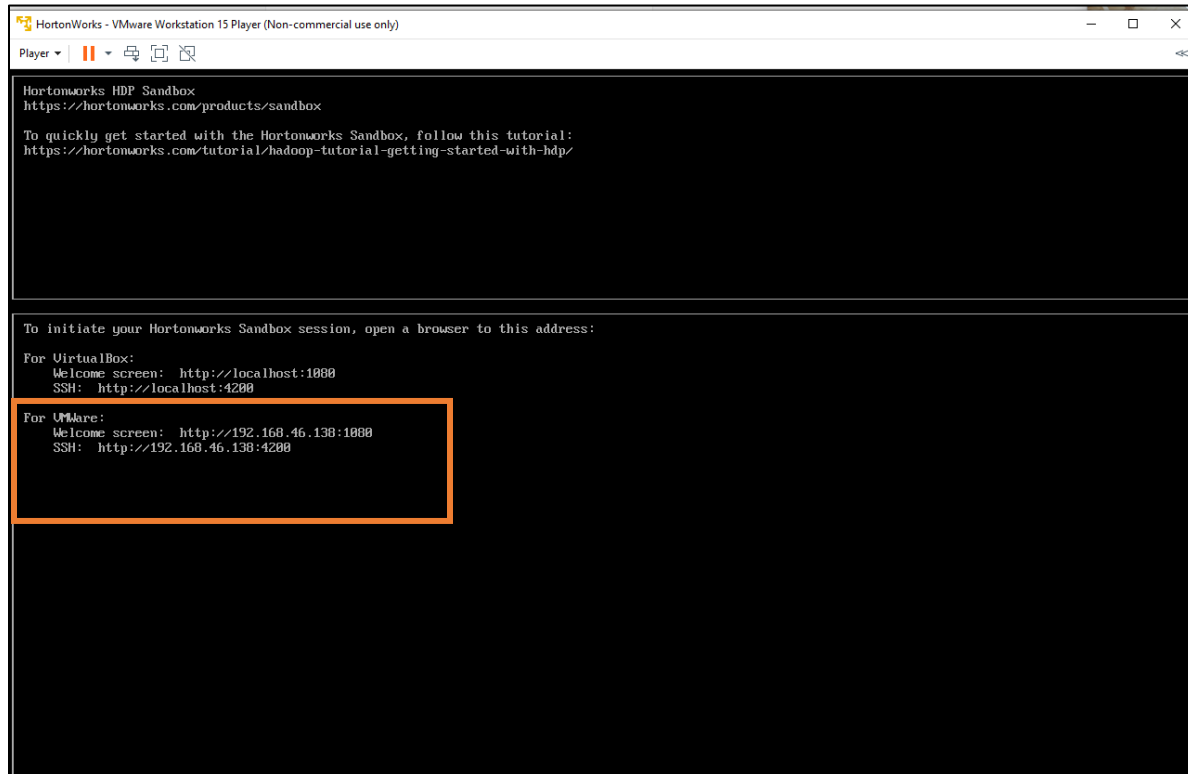
We have connected cityid from project table to location table, to form a many to many relationships and establish relationship. This location table helps business to understand which countries have maximum projects within Kickstarter or gross earnings by each country. The team can also analyze the duration of the projects to complete. Is it because of long duration the projects or insufficient backers the projects are not successful? We have performed analysis and created data visualizations to understand all these issues for businesses using MS PowerBI

HSBOX Installation

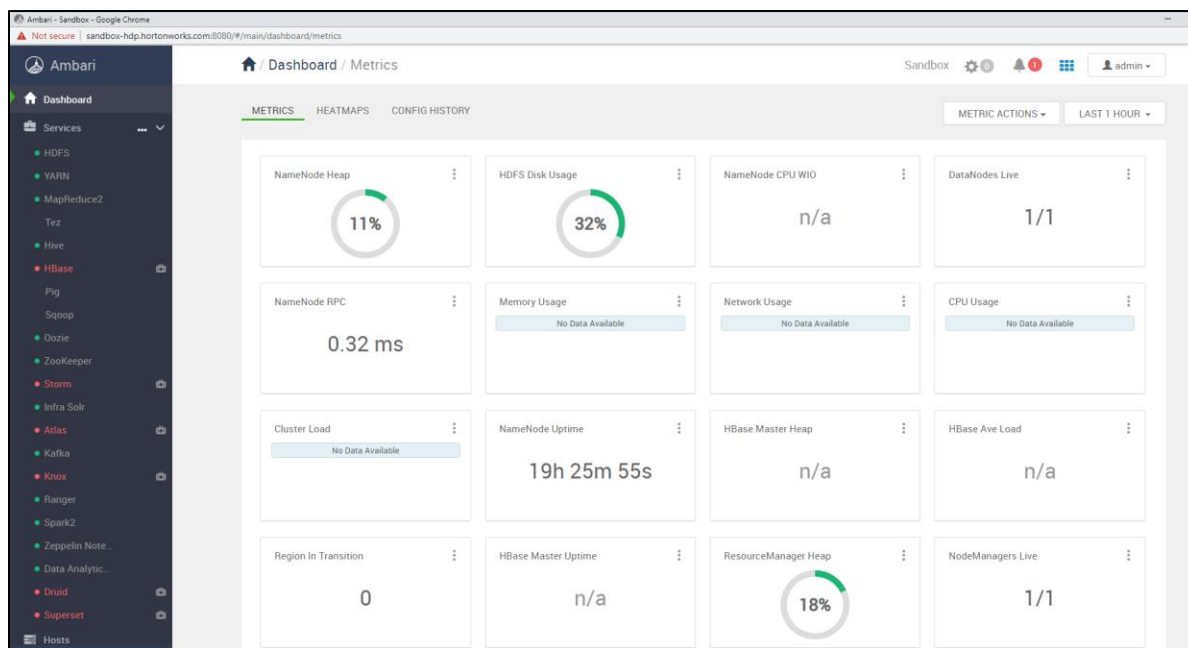
Downloaded and installed VMWare workstation player from the web. After installing player, we downloaded Horton Works Sandbox from Cloudera website. Created a virtual machine in VMware workstation and imported the vdc file downloaded from cloudera.



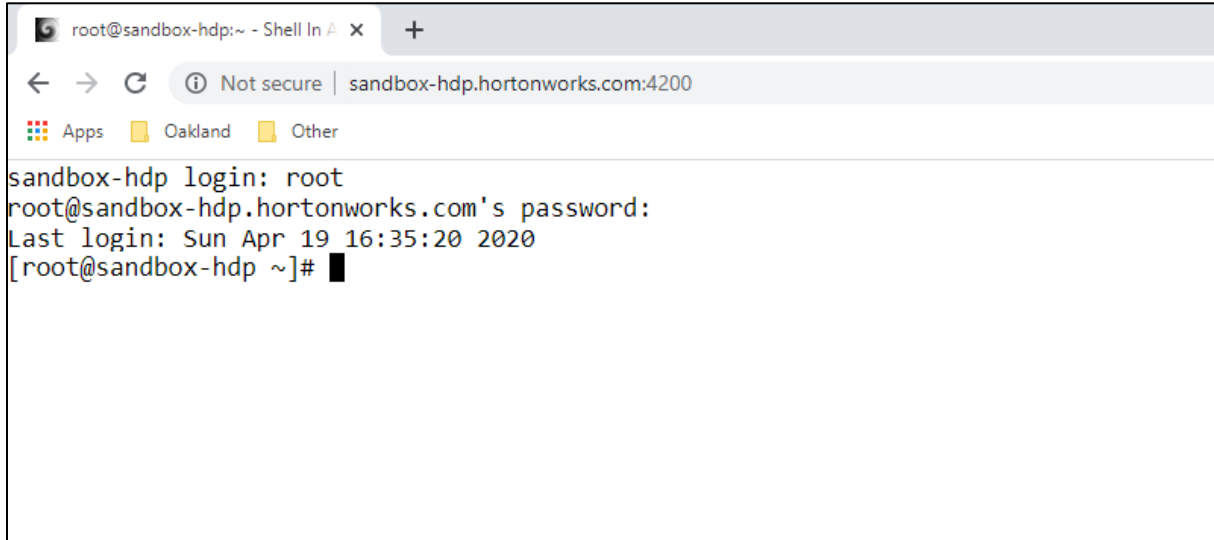
Once the files are imported and the virtual machine fully boots up, we will see a similar screen as below.



We can enter the details for VMware in the browser and will be connected to the web access of Horton Works Sandbox (Ambari).



Ambari is the web access view for Horton Works Sandbox. We need to ensure that all the services are up (**YARN, HIVE, TEZ and HDFS**) in running in before we navigate to command prompt. We can add new users from Ambari and provide required access. Command prompt can be accessed at port 4200 of Horton works sandbox. We can login to command prompt and initialize project steps

A screenshot of a web browser window. The address bar shows 'sandbox-hdp.hortonworks.com:4200' with a 'Not secure' warning. The page content displays a terminal session for logging into the sandbox. The text shown is: 'sandbox-hdp login: root', 'root@sandbox-hdp.hortonworks.com's password:', 'Last login: Sun Apr 19 16:35:20 2020', and '[root@sandbox-hdp ~]#'. There is a black cursor at the end of the last line.

```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Sun Apr 19 16:35:20 2020
[root@sandbox-hdp ~]#
```

Data Import to HDFS and Hive - Horton Works Sandbox

We exported data from local drive to GitHub from where it was then imported to the local drive of Horton Works Sandbox using command prompt (Shell in a Box).

Data from the source file was imported as different .csv files and was saved to a directory in the local drive(/user) of HW Sandbox

```
root@sandbox-hdp ~ Shell In X
Not secure | sandbox-hdp.hortonworks.com:4200
Apps Oakland Other HDP links
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Fri Apr 3 12:28:28 2020
[root@sandbox-hdp ~]# hdfs dfs -ls /user
Found 17 items
[root@sandbox-hdp ~]# wget https://raw.githubusercontent.com/asjustin83/KickStarter/master/Category.csv
--2020-04-03 12:42:52-- https://raw.githubusercontent.com/asjustin83/KickStarter/master/Category.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.200.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.200.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3021 (3.0K) [text/plain]
Saving to: 'Category.csv'

100%[=====] 3,021  --K/s  in 0s

2020-04-03 12:42:53 (11.5 MB/s) - 'Category.csv' saved [3021/3021]

[root@sandbox-hdp ~]# wget https://raw.githubusercontent.com/asjustin83/KickStarter/master/Creator.csv
--2020-04-03 12:43:06-- https://raw.githubusercontent.com/asjustin83/KickStarter/master/Creator.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.248.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.248.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1048395 (1024K) [text/plain]
Saving to: 'Creator.csv'
```

- Checked if the all the files were imported and verified details in the file

```
root@sandbox-hdp ~ Shell In X
Not secure | sandbox-hdp.hortonworks.com:4200
Apps Oakland Other HDP links
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg Category.csv Creator.csv Currency.csv Location.csv Project.csv
[root@sandbox-hdp ~]# ls -l
total 5964
-rw-r--r-- 1 root root 3302 May 31 2018 anaconda-ks.cfg
-rw-r--r-- 1 root root 3021 Apr 3 12:42 Category.csv
-rw-r--r-- 1 root root 1048395 Apr 3 12:43 Creator.csv
-rw-r--r-- 1 root root 511 Apr 3 12:43 Currency.csv
-rw-r--r-- 1 root root 150191 Apr 3 12:43 Location.csv
-rw-r--r-- 1 root root 4892394 Apr 3 12:43 Project.csv
[root@sandbox-hdp ~]# cat Category.csv
Category,SubCategory
Art,Art
Film & Video,Short Film
Fashion,Fashion
Publishing,Poetry
Publishing,Nonfiction
Film & Video,Narrative Film
Design,Product Design
Technology,Hardware
Photography,Photography
Technology,Technology
Publishing,Art Book
Art,Performance Art
Design,Design
Film & Video,Documentary
Food,Food
Technology,Open Software
Music,Indie Rock
Journalism,Journalism
Publishing,Radio & Podcast
Film & Video,Film & Video
Games,Tabletop Games
Art,Illustration
Games,Video Games
Art,Public Art
Music,Country & Folk
Theater,Theater
Crafts,Crafts
Music,Music
```

- Created directories in HDFS and moved data from local drive to HDFS directories.

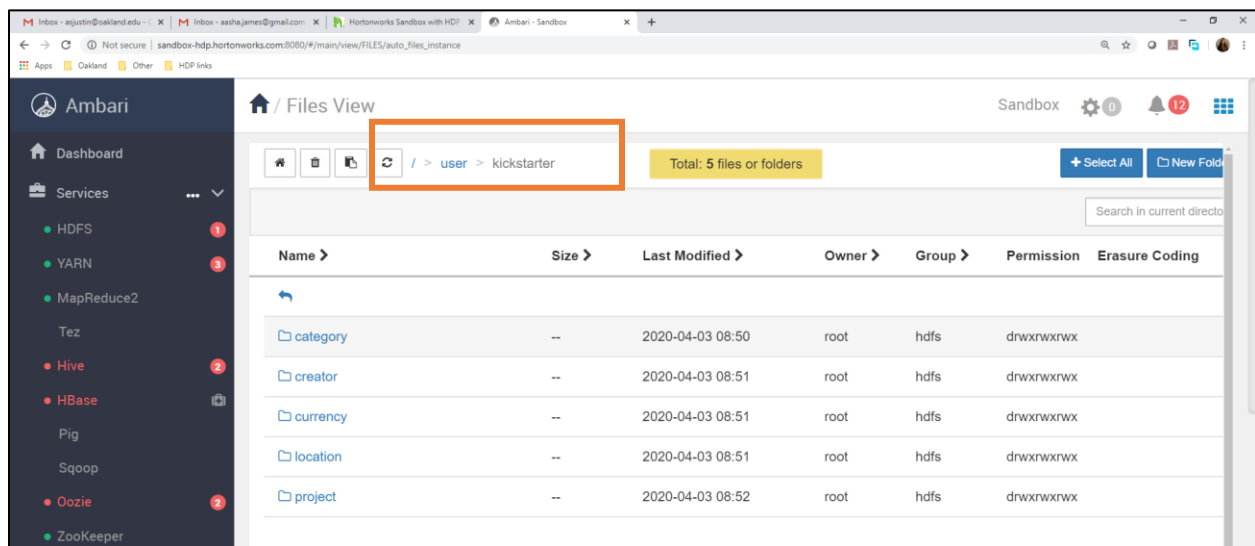
```

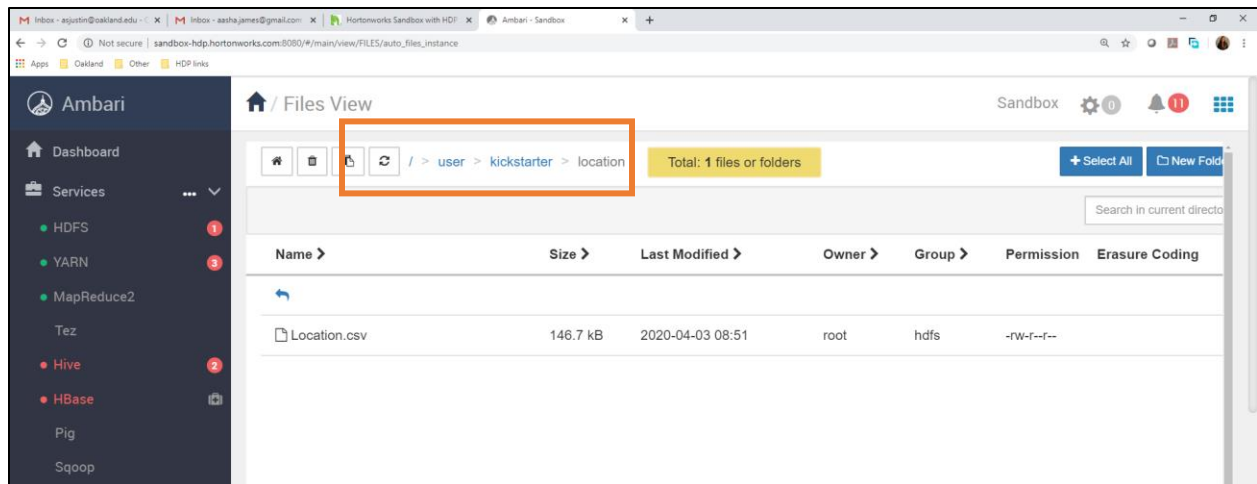
root@sandbox-hdp:~ - Shell In A x Manage Files on HDFS with the C x +
Not secure | sandbox-hdp.hortonworks.com:4200
Apps Oakland Other HDP links

[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter
mkdir: '/user/kickstarter': File exists
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter/category
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter/currency
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter/creator
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter/location
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/kickstarter/project
[root@sandbox-hdp ~]# hdfs dfs -put Category.csv /user/kickstarter/category
[root@sandbox-hdp ~]# hdfs dfs -put Currency.csv /user/kickstarter/currency
[root@sandbox-hdp ~]# hdfs dfs -put Creator.csv /user/kickstarter/creator
[root@sandbox-hdp ~]# hdfs dfs -put Location.csv /user/kickstarter/location
[root@sandbox-hdp ~]# hdfs dfs -put Project.csv /user/kickstarter/project
[root@sandbox-hdp ~]# hdfs dfs -ls /user/kickstarter
Found 5 items
drwxr-xr-x - root hdfs 0 2020-04-01 16:19 /user/kickstarter/category
drwxr-xr-x - root hdfs 0 2020-04-01 16:23 /user/kickstarter/creator
drwxr-xr-x - root hdfs 0 2020-04-01 16:23 /user/kickstarter/currency
drwxr-xr-x - root hdfs 0 2020-04-01 16:24 /user/kickstarter/location
drwxr-xr-x - root hdfs 0 2020-04-01 16:25 /user/kickstarter/project
[root@sandbox-hdp ~]# exit
logout
Connection to sandbox-hdp.hortonworks.com closed.
Session closed.

```

- Verified data uploaded on HDFS using the WebView (Ambari)





- We can create tables in hive using the web access of Horton Sandbox from Data Studio Analytics by importing the csv files uploaded to HDFS. But for our project we used command prompt to create tables and upload data to these tables from HDFS directories.
 - Created external tables in hive and imported data from HDFS
 - Created ORC tables and moved data from external tables to these ORC tables and delete external tables.
 - Verified if these tables were successfully created by writing queries in command prompt and in the web access


```
root@sandbox-hdp-9 - Shell in %  
← → ↺ Ⓜ Not secure | sandbox-hdp.hortonworks.com:4200  
Apps Oakland Other HDP links  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20200405143601_668b5510-0d01-4927-a9a9-eba8b668a7f4); Time taken: 0.014 seconds  
INFO : OK  
-----+  
| database_name |  
-----+  
| default      |  
| foodmart     |  
| information_schema |  
| kickstarter  |  
| kickstarterprj |  
| mytestdb    |  
| sys         |  
-----+  
7 rows selected (0.126 seconds)  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> use kickstatrterprj;  
Error: Error while compiling statement: FAILED: SemanticException [Error 10072]: Database does not exist: kickstatrterprj (state=42000,code=10072)  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> use kickstarterprj;  
INFO : Compiling command(queryId=hive_20200405143634_ccec7af4-ce86-465a-8732-6b9cd3844a75): use kickstarterprj  
INFO : Semantic Analysis Completed (retrial = false)  
INFO : Returning Hive schema: Schema(fieldsSchemas:null, properties:null)  
INFO : Completed compiling command(queryId=hive_20200405143634_ccec7af4-ce86-465a-8732-6b9cd3844a75); Time taken: 0.048 seconds  
INFO : Executing command(queryId=hive_20200405143634_ccec7af4-ce86-465a-8732-6b9cd3844a75): use kickstarterprj  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20200405143634_ccec7af4-ce86-465a-8732-6b9cd3844a75); Time taken: 0.008 seconds  
INFO : OK  
No rows affected (0.063 seconds)  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> show tables;  
INFO : Compiling command(queryId=hive_20200405143641_419dab17-960f-4a15-9fae-3acb13465c9c): show tables  
INFO : Semantic Analysis Completed (retrial = false)  
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=tab_name, type:string, comment:from deserializer)], properties:null)  
INFO : Completed compiling command(queryId=hive_20200405143641_419dab17-960f-4a15-9fae-3acb13465c9c); Time taken: 0.022 seconds  
INFO : Executing command(queryId=hive_20200405143641_419dab17-960f-4a15-9fae-3acb13465c9c): show tables  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20200405143641_419dab17-960f-4a15-9fae-3acb13465c9c); Time taken: 0.011 seconds  
INFO : OK  
-----+  
| tab name |  
-----+  
| category |  
| creator  |  
| currency |  
| location |  
| project  |  
-----+  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1>
```

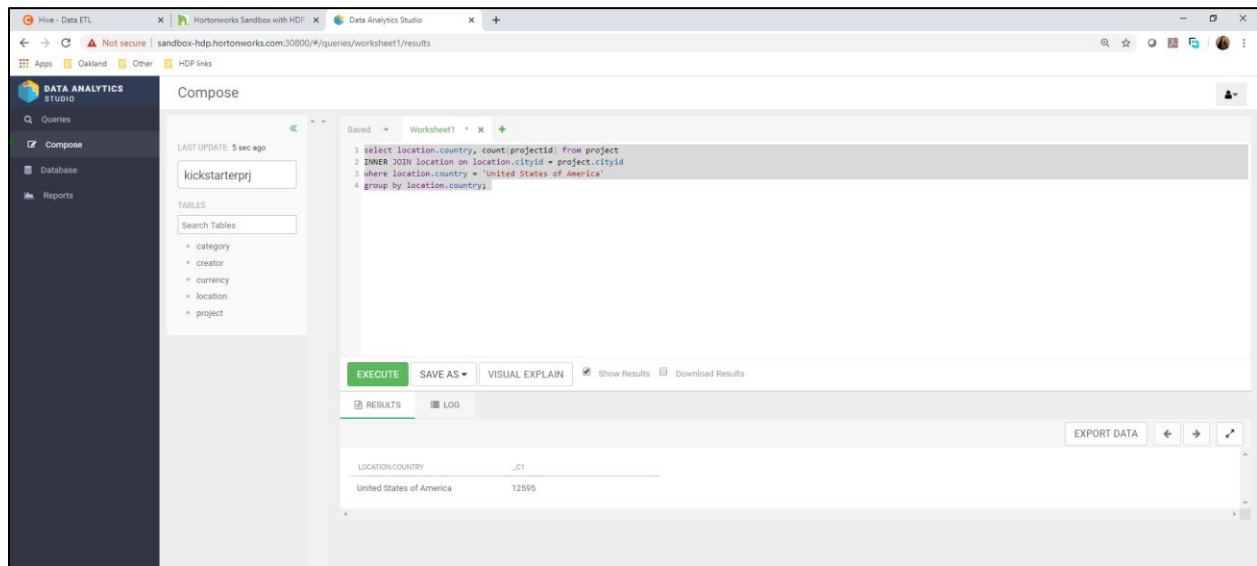
Executed queries both in command prompt and web view to check if the output is identical.

```
.> select location.country, count(projectid) from project
.> INNER JOIN location on location.cityid = project.cityid
.>
.> where location.country = 'United States of America'
.> group by location.country;
```

```
INFO : INPUT_FILES_Map_3: 1
INFO : RAW_INPUT_SPLITS_Map_1: 1
INFO : RAW_INPUT_SPLITS_Map_3: 1
INFO : Completed executing command(queryId=hive_20200405144344_e8b19200-5630-4411-b0ba-a17c3977ba46); Time taken: 2123.885 seconds

+-----+-----+
| location.country | c1 |
+-----+-----+
| United States of America | 12595 |
+-----+-----+

1 row selected (2124.279 seconds)----+
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> █
```



From the query executed above we can see that we have 12595 projects for United States of America. Similarly, we can execute multiple queries in Data Analytics Studio from the tables created in Hive.

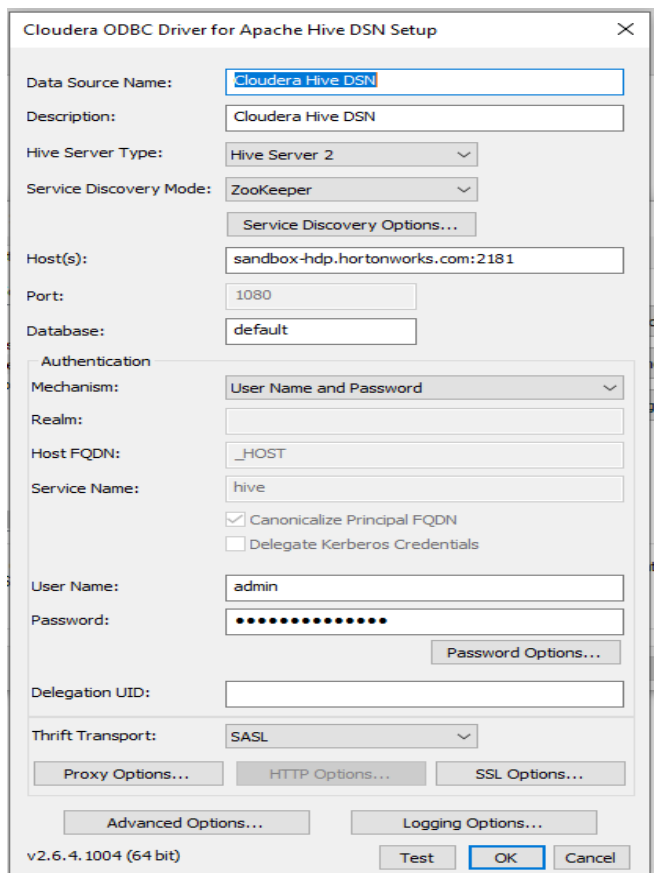
Data Visualization –

Data visualization can be performed within Horton Works Sandbox using Data Analytics Studio using the data from Hive tables. But for advanced visualization techniques we chose Microsoft Power BI in our project.

Installed PowerBI desktop version from - <https://powerbi.microsoft.com/en-us/downloads>. We initialized a connection between Horton Works and PowerBI using ODBC.

To enable the connection, we need to add ODBC as a connector to PowerBI.

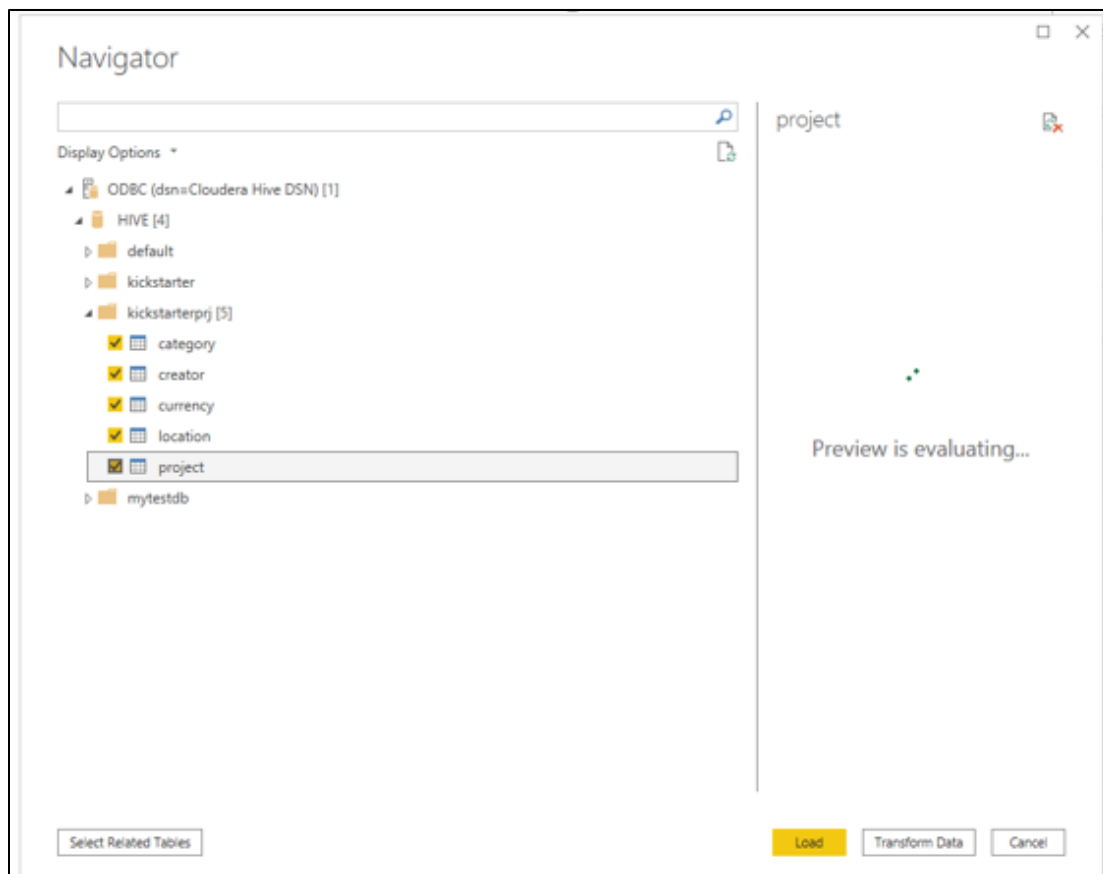
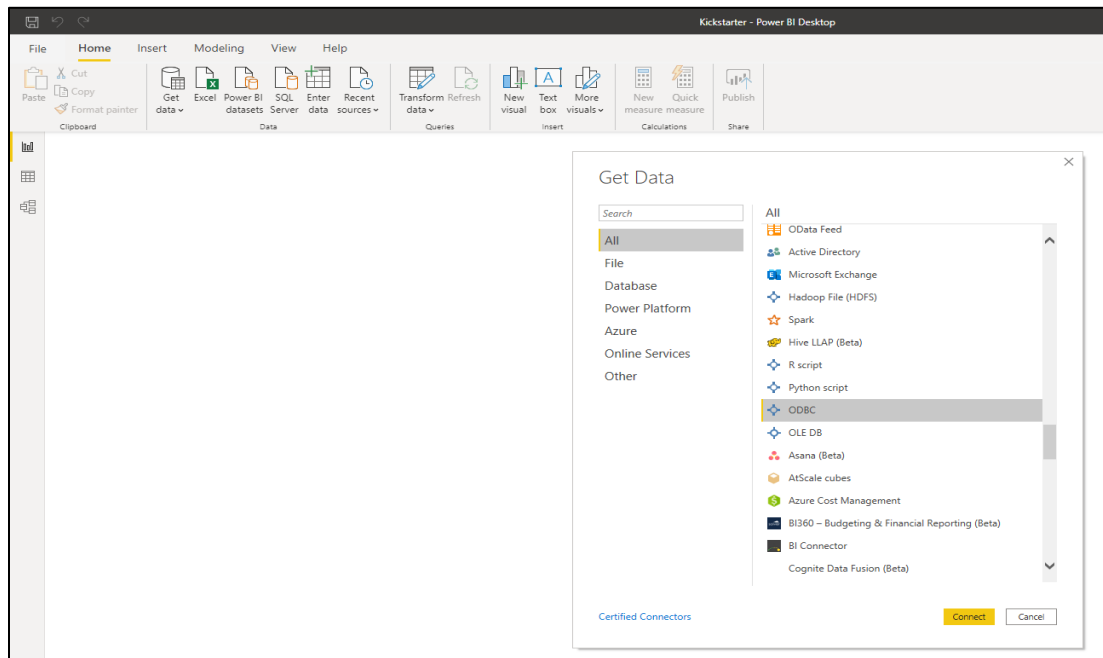
Installed ODBC driver for Horton Works Sandbox from Cloudera website and configured the same to connect to PowerBI.



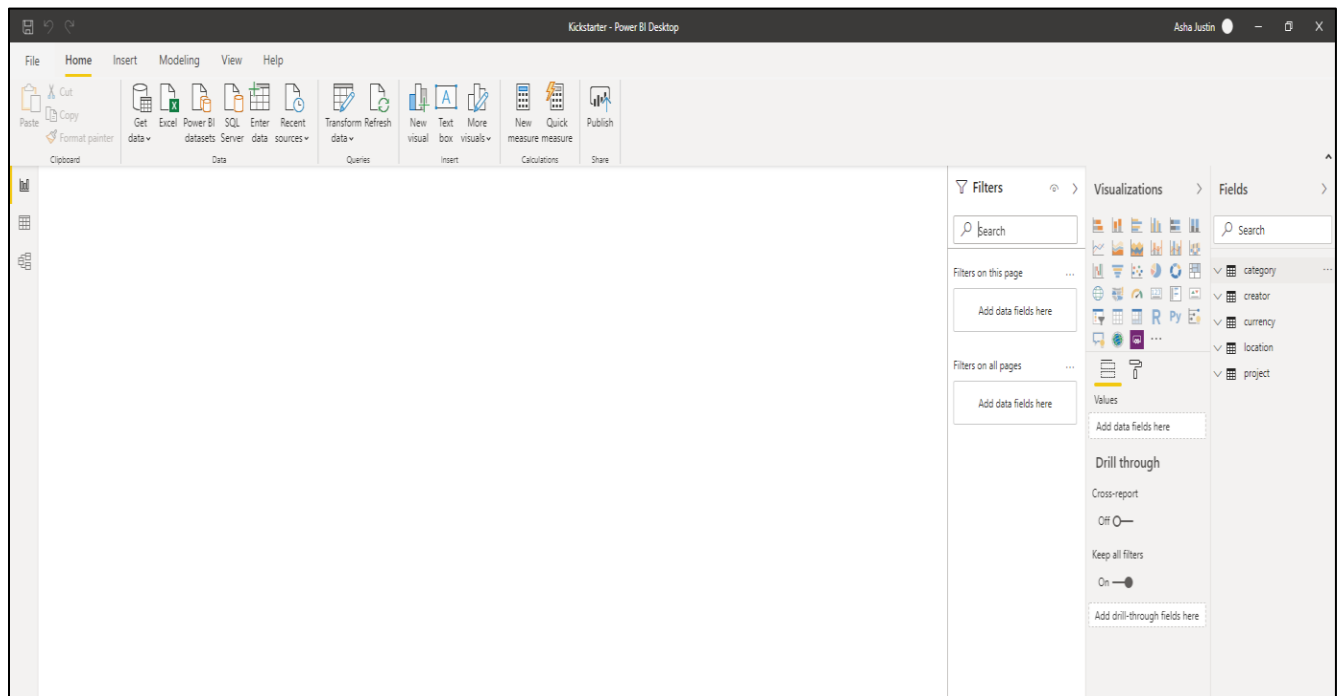
The screenshot shows the 'Cloudera ODBC Driver for Apache Hive DSN Setup' dialog box. The fields are configured as follows:

- Data Source Name: Cloudera Hive DSN
- Description: Cloudera Hive DSN
- Hive Server Type: Hive Server 2
- Service Discovery Mode: ZooKeeper
- Host(s): sandbox-hdp.hortonworks.com:2181
- Port: 1080
- Database: default
- Authentication Mechanism: User Name and Password
- Realm: (empty)
- Host FQDN: _HOST
- Service Name: hive
- Canonicalize Principal FQDN: ☒
- Delegate Kerberos Credentials: ☐
- User Name: admin
- Password: (masked with dots)
- Delegation UID: (empty)
- Thrift Transport: SASL
- Proxy Options..., HTTP Options..., SSL Options..., Advanced Options..., and Logging Options... buttons are visible.
- Version: v2.6.4.1004 (64 bit)
- Buttons: Test, OK, Cancel

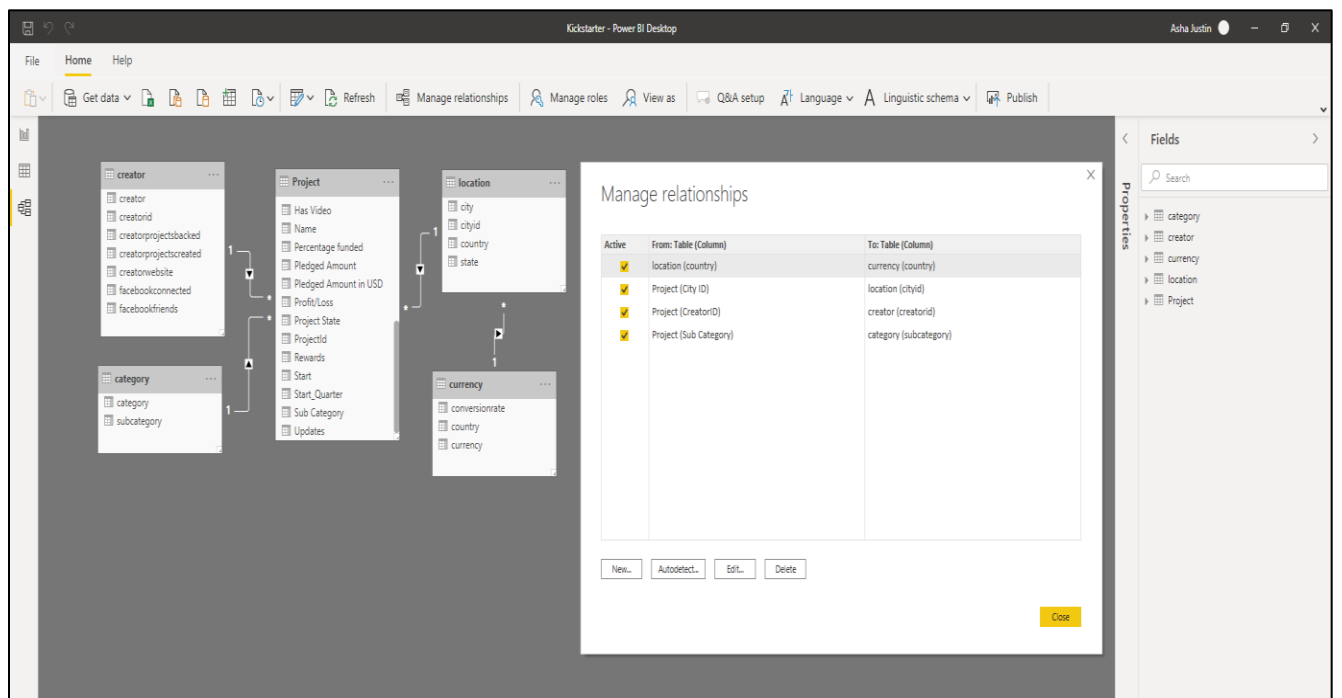
Open PowerBI and click on Get data and select ODBC. Once the connection is enabled, we can view all the tables and databases that are created in Hive.



Click on Load and all the tables from Hive will be loaded to PowerBI

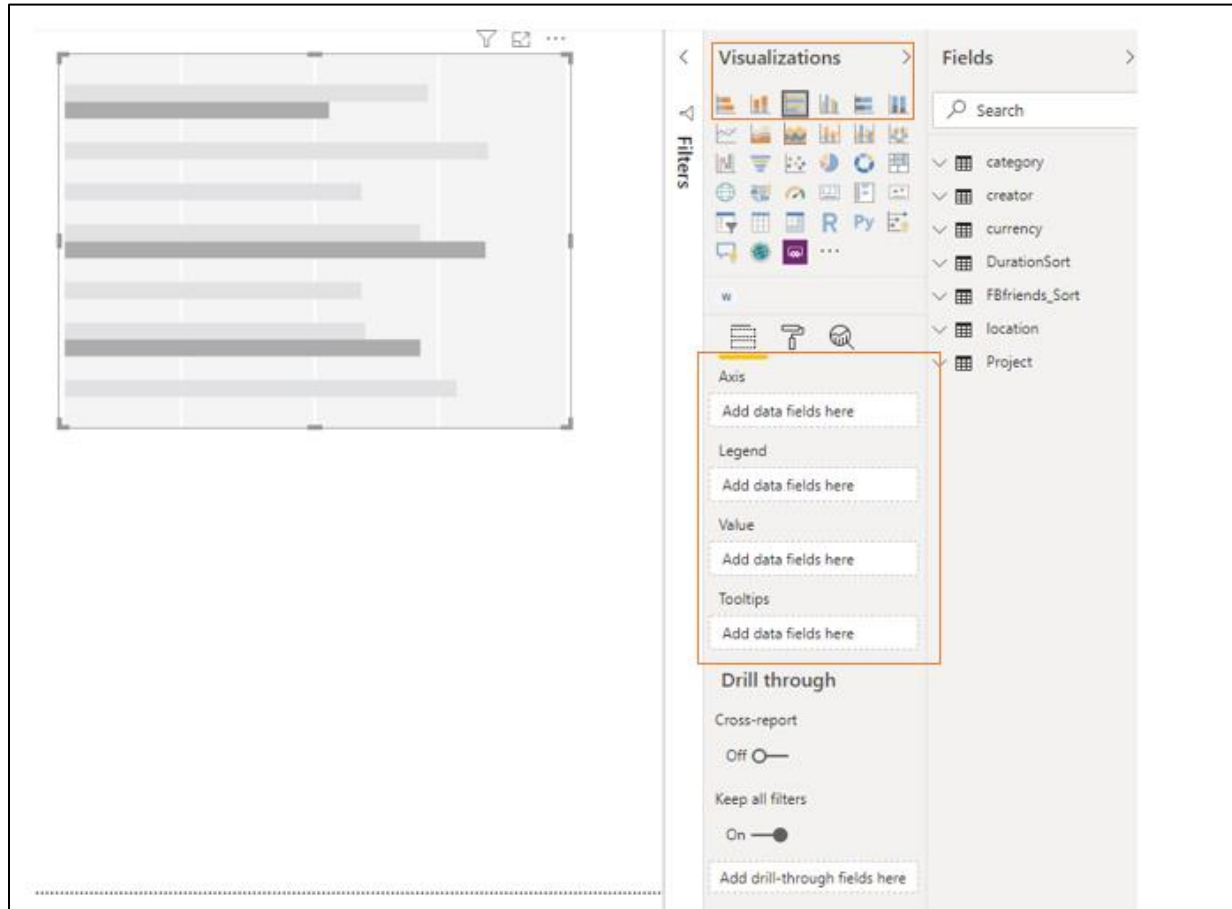


After the tables are uploaded, we need to manage relationship between the tables and modify the primary key and foreign key details (if required) for all the tables in the database.



Once the tables are managed as per the relationship, we can now create different charts and analyze management requirements as suggested.

To create a graph, we can select a chart from the visualization ribbon and add details in Legend, Axis and Values fields based on graph type selected.



We have selected different visualizations available in Powerbi to create a report which will analyze all the data available in the tables.

We also created groups in PowerBI to analyze numerical data better

Groups

Name: FBfriends_Group Field: facebookfriends

Group type: List

Ungrouped values

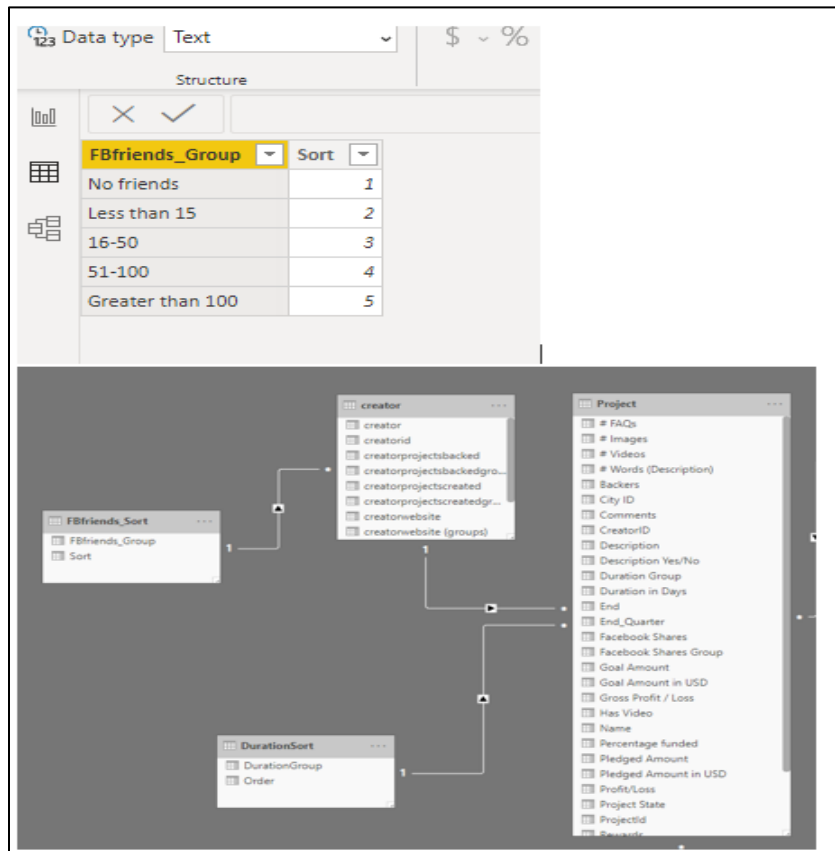
Groups and members

- ▶ 16-50
- ▶ 51-100
- ▶ Greater than 100
- ▶ Less than 15
- ▶ No Friends
- ▶ Other

☒ Include Other group

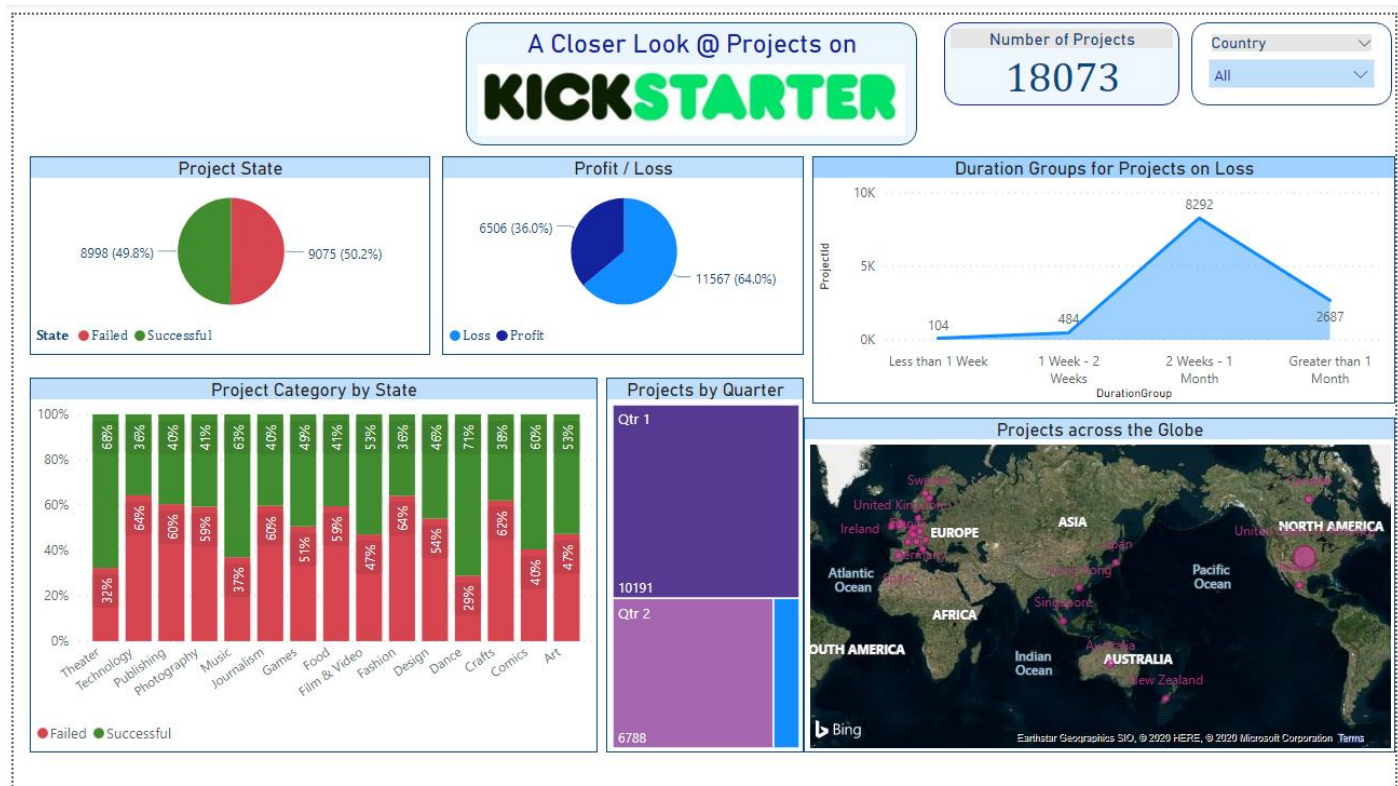
Group Ungroup OK Cancel

Created separate tables to manage order of the variables which follow an order so that the graphs can updated accordingly. Created relationship with the parent table



Overview

This page gives an overall view about the company Kickstarter. We can select country details from the slicer and the graphs in the sheet will refresh accordingly. We can see number of projects available on the database are 18073 with the status as failed (8998) or successful (9075). We can also view overall status about profit or loss in these projects. We can analyze that Dance and Theatre categories see a good success rate as compared to other project categories. Projects which are in loss take a longer time to complete which can be one of the reasons for their loss. More projects start in Quarter 1 and 2 than the later quarters of the year. Most number of projects available on Kickstarter are from United States of America.



Top 10

This page will give details about Top 10 values by each category. We can filter the state of the project as successful or failed and can analyze each of the categories available on the report.

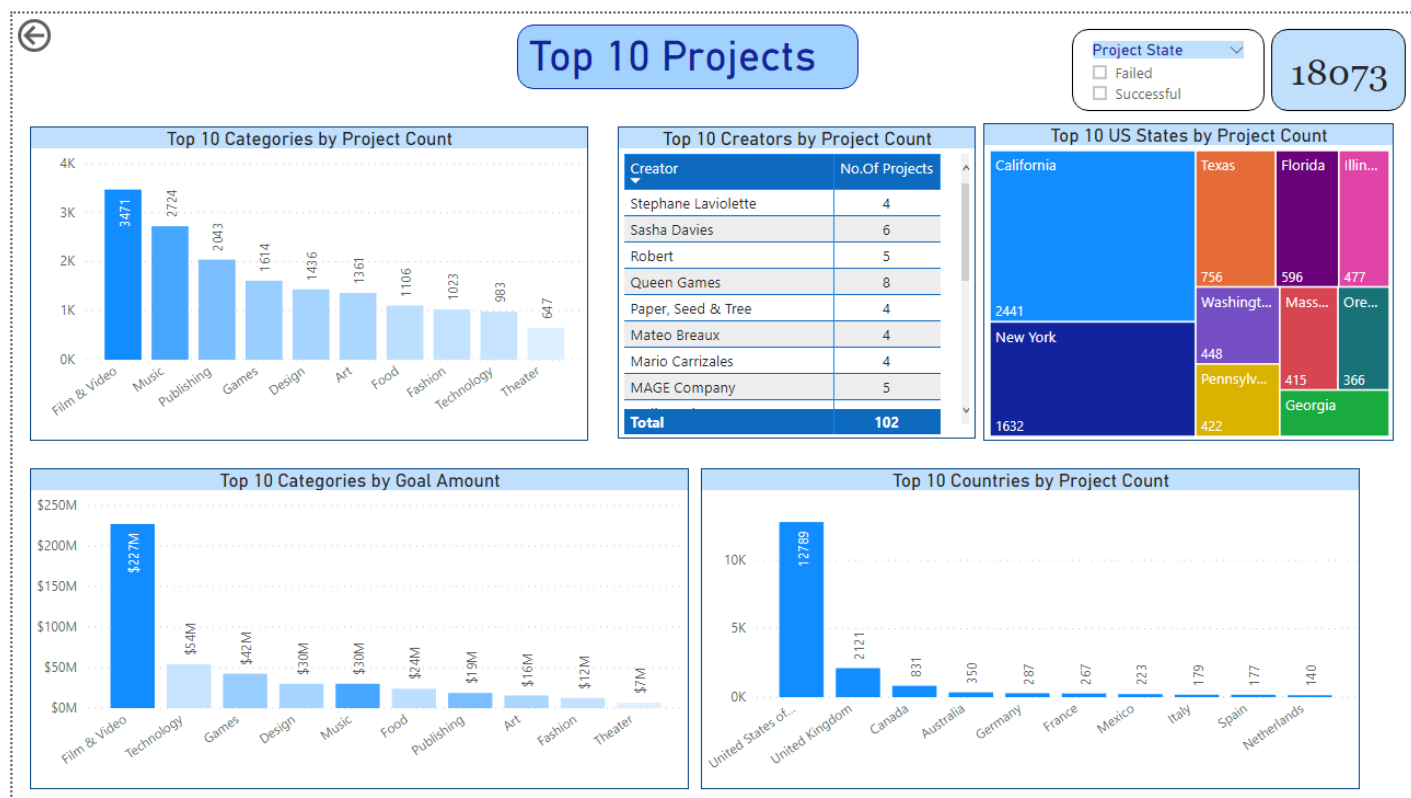
Top successful category by project count – Film & Video (1837)

Top successful creator by project count – Benjamin Hennessey (13)

Top successful country by project count – United States of America (12789)

Top successful category by Goal Amount - Film & Video(\$21.1M)

California is the top state in United States of America by number of projects invested at Kickstarter



Creators and Backers

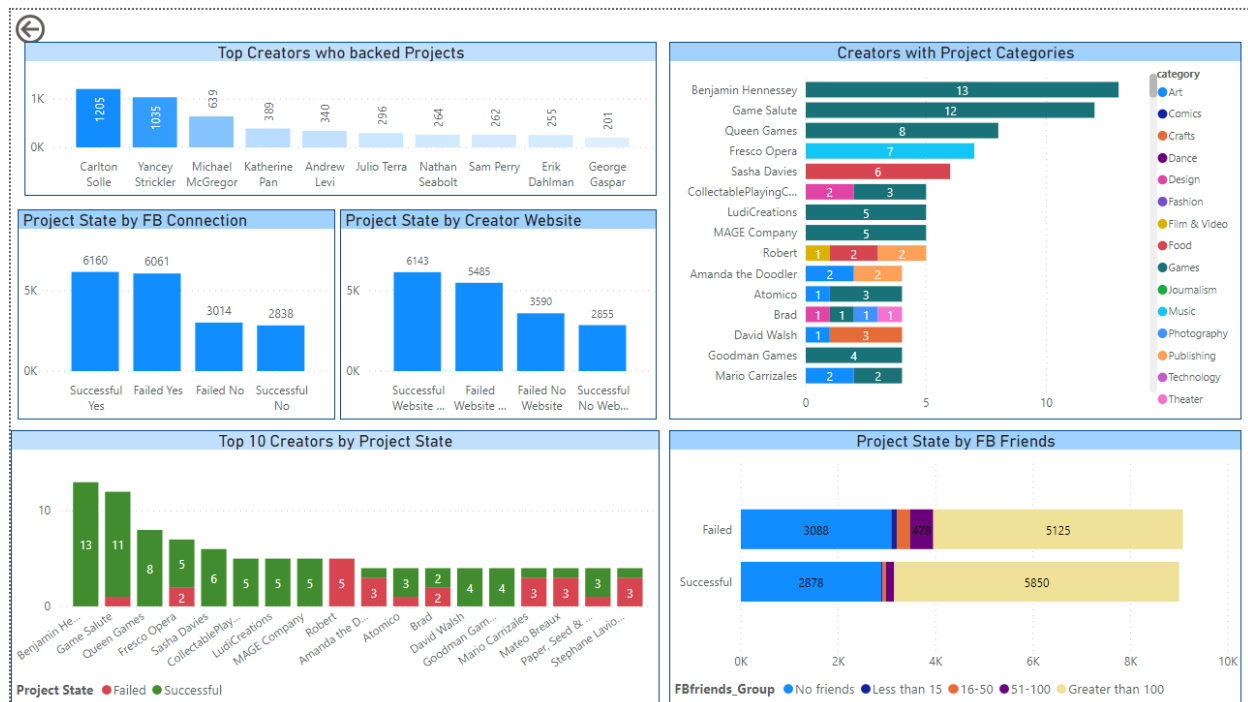
This page will analyze how creators and backers impact the success or failure of the project. We can find who are the creators backing other projects, creators creating projects in different categories.

We can also analyze having a Facebook connection or having a website impact the success rate of the project. From the graph below we can suggest that having Facebook connection does not affect much but having a website affects the success of a project. (6143 vs 5485)

But more Facebook friends a creator has more are the chances of success. Reason maybe he shares the details on Facebook which might attract more backers to fund the project and creator reaches the Goal Amount required to create his project.

From the graph below we can see that all projects for Robert have failed whereas Benjamin has all projects successful. We can also see a trend of creators creating different types of project. Brad has 4 projects and all in different categories

We can click on any of the views and get a detailed view about other parameters related .

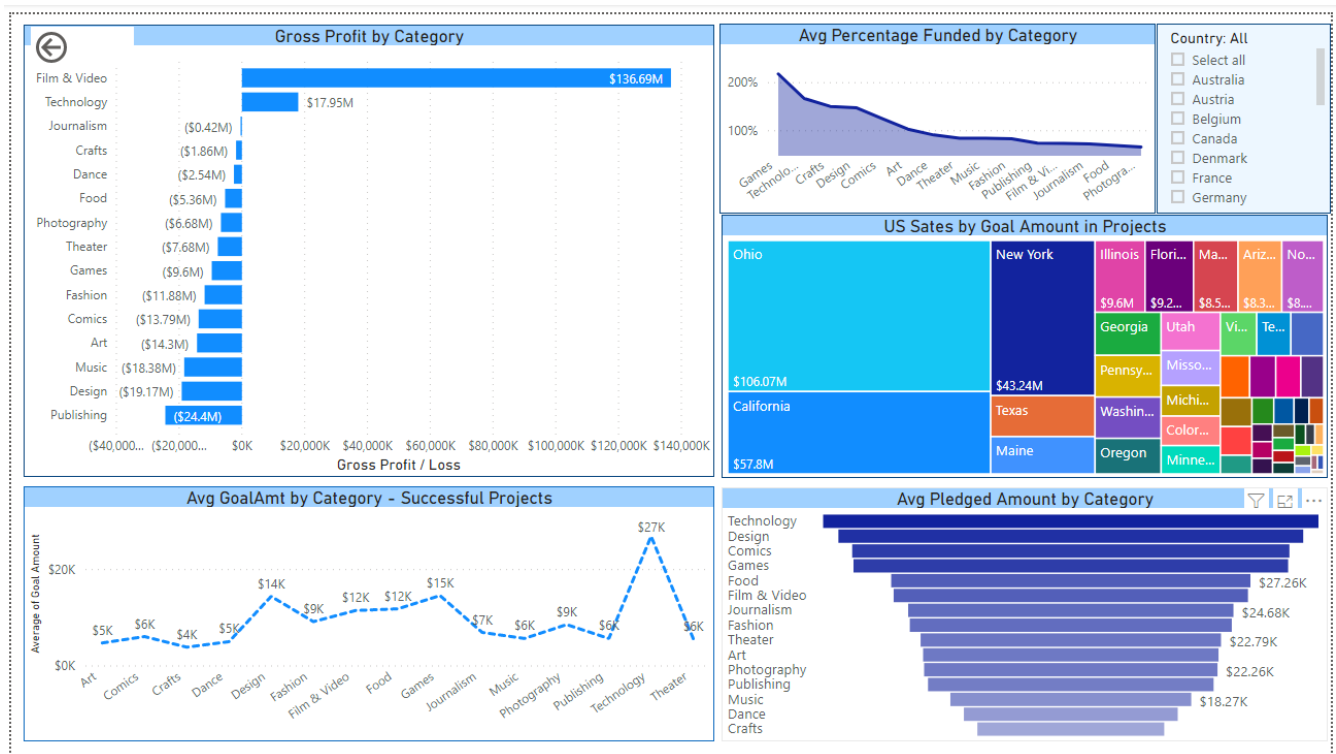


Revenue Analysis

This page will analyze the health of the company Kickstarter based on the revenue from projects. From the graph below we can see that only categories that are making average gross profit are Film & Video and Technology. Backers are more inclined in funding Games and Technology as compared to food or photography. We can also see the trend for all these points by country also by clicking on the slicer (if required).

Ohio is the US state which has maximum Goal amount to be invested in projects. Technology and Design projects have maximum average pledged amount at Kickstarter.

We can also analyze the average goal amount required to create different project types. Technology projects require more money (\$27K) whereas art projects are the easiest to be funded.



Challenges

- *Writing MDX queries to analyze data in Visual studio cubes*
- *Analyze & Predict data using different techniques*
- *Installation and Configuration of Horton Works Sandbox*
- *Enhancing Data Visualization Techniques in PowerBI*

Conclusion

- *Most successful projects by Country or Category – US & Film & Video*
- *Average Goal Amount for a successful project by Category - \$9.6K*
- *Creators with most successful projects – Benjamin Hennessey (13 projects)*
- *Having Facebook connection /website effect success of a project – FB connection ~ No
Creator Website ~ Yes*
- *Which categories have projects with maximum profit – Film & Video*
- *Who are the people creating projects in different categories – Playing Cards*
- *What are the reasons for success / failure of the projects – More duration, Less backers, No website*

After analyzing the data, we can conclude that an investor can invest in a Film & Video project to make a bigger gross profit. Having a creator website will attract more backers and the project can be funded easily. We can monitor time taken to complete the project and ensure that it is completed before the agreed timelines.

Lessons Learned

- *Learned to create cubes & reports with MS Visual Studio*
- *Understood to execute MDX queries.*
- *Deep understanding of data and details related to it*
- *Data Analysis using cubes and data mining techniques*
- *Dealing with big data*
- *Working with Hive*
- *Extensive data visualization and data analysis using Microsoft PowerBi*
- *Data prediction using above techniques*

References

- <https://www.cloudera.com/tutorials/learning-the-ropes-of-the-hdp-sandbox.html>
- <https://www.cloudera.com/tutorials/getting-started-with-hdp-sandbox.html>
- <https://powerbi.microsoft.com/en-us/>
- <https://towardsdatascience.com/using-machine-learning-to-predict-kickstarter-success-e371ab56a743>
- <https://docs.microsoft.com/en-us/sql/mdx/mdx-function-reference-mdx?view=sql-server-ver15>
- <https://community.cloudera.com/t5/Community-Articles/Visualizing-Hive-Data-Using-Microsoft-Power-BI/ta-p/247769>