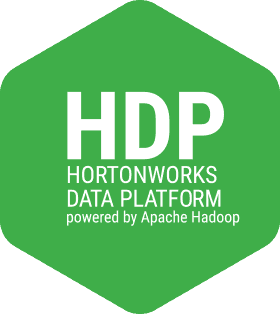
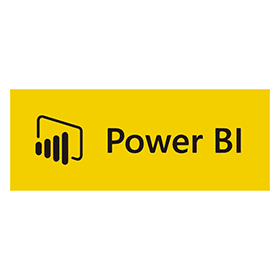
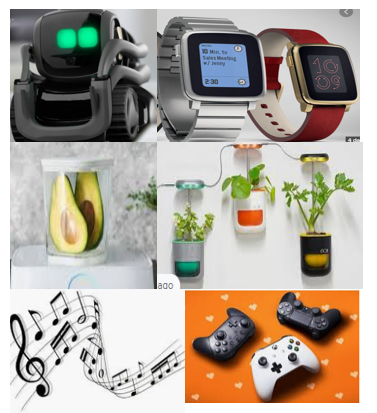
**Advance Databases and Big Data Management**

**A Closer Lo****k @Projects on**

**KICKSTARTER**

Dream Maker or Promise Breaker



Prof. Vijayan Sugumaran

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.

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.

**BUSINESS USES FROM THE PROJECT**

**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**

**Creator**

**CreatorID (PK)**

Name

Facebook Connected

Facebook Friends

Creator Website

**ProjectID (PK)**

Name

URL

Project State

Description

Category

**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

**Date**

**Location**

**DateID (PK)**

StartDate

EndDate

Month

Year

Quarter

Quarter

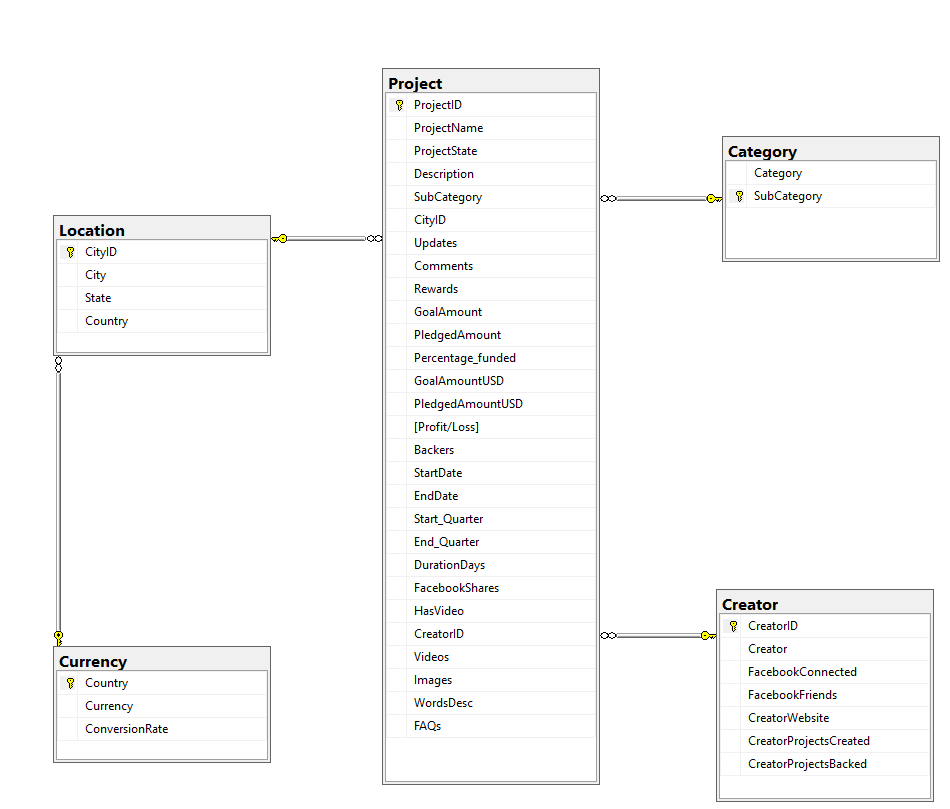
**LocationID (PK)**

City

State

Country

**DATAWARE HOUSE MODEL – SQL**



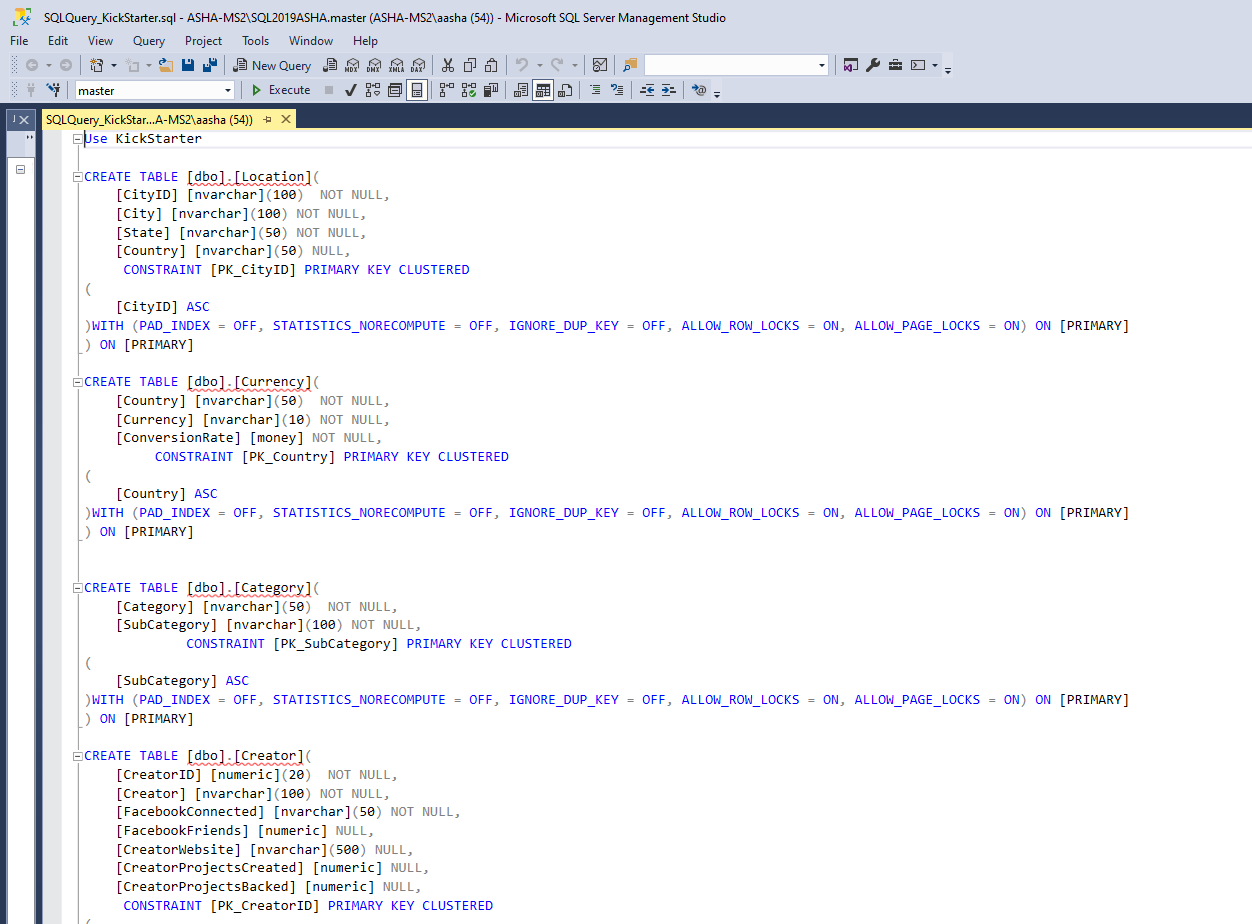
**MODEL 1 – Visual Studio**

**MODEL**

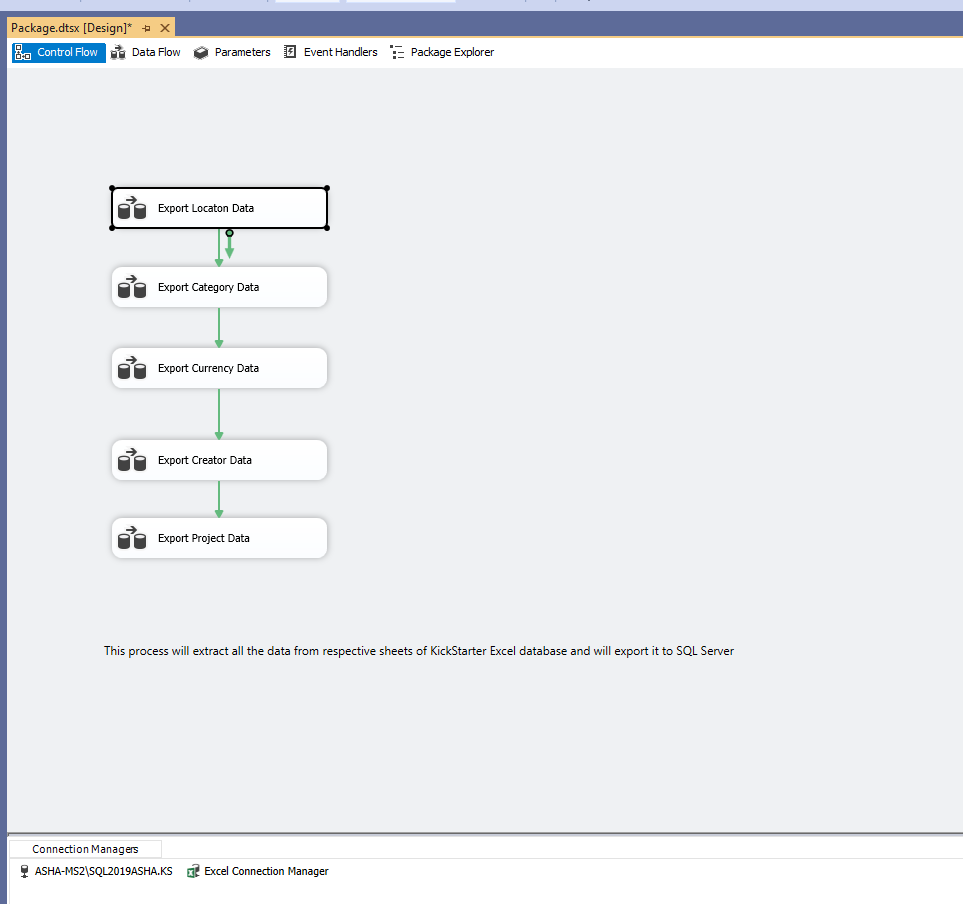
Model was built in Visual Studio by creating a data warehouse, and extracting data into SQL using Integration Services Project. After the data is loaded into SQL, we created cubes and reports to analyze the data. This analysis will help the management team understand which countries or categories invest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

**Extraction Transformation Load**

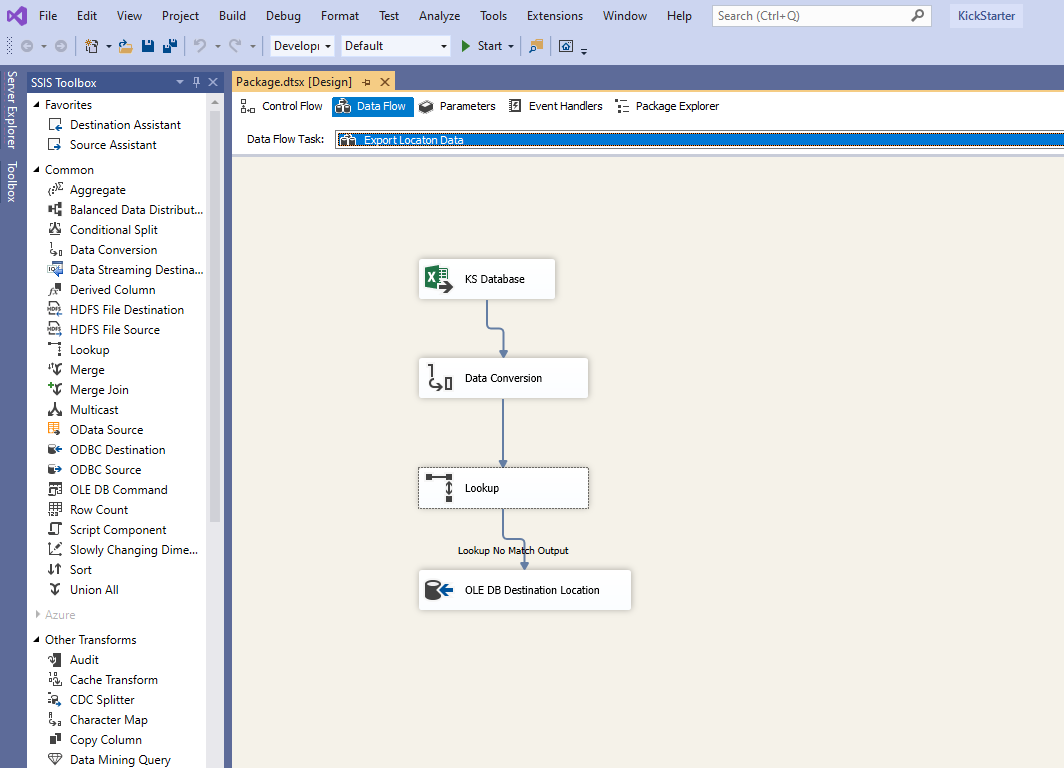
Created tables (Destination) in SQL where the data extracted from the excel file (data source) can be loaded.



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.

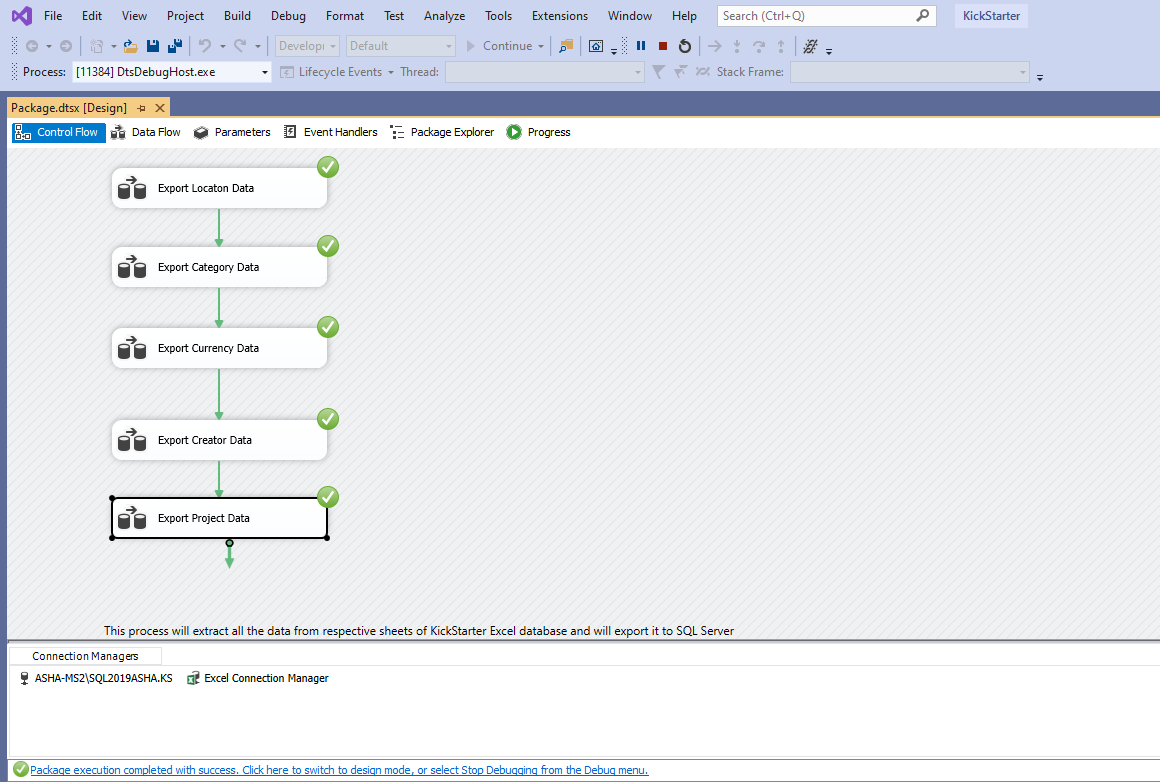


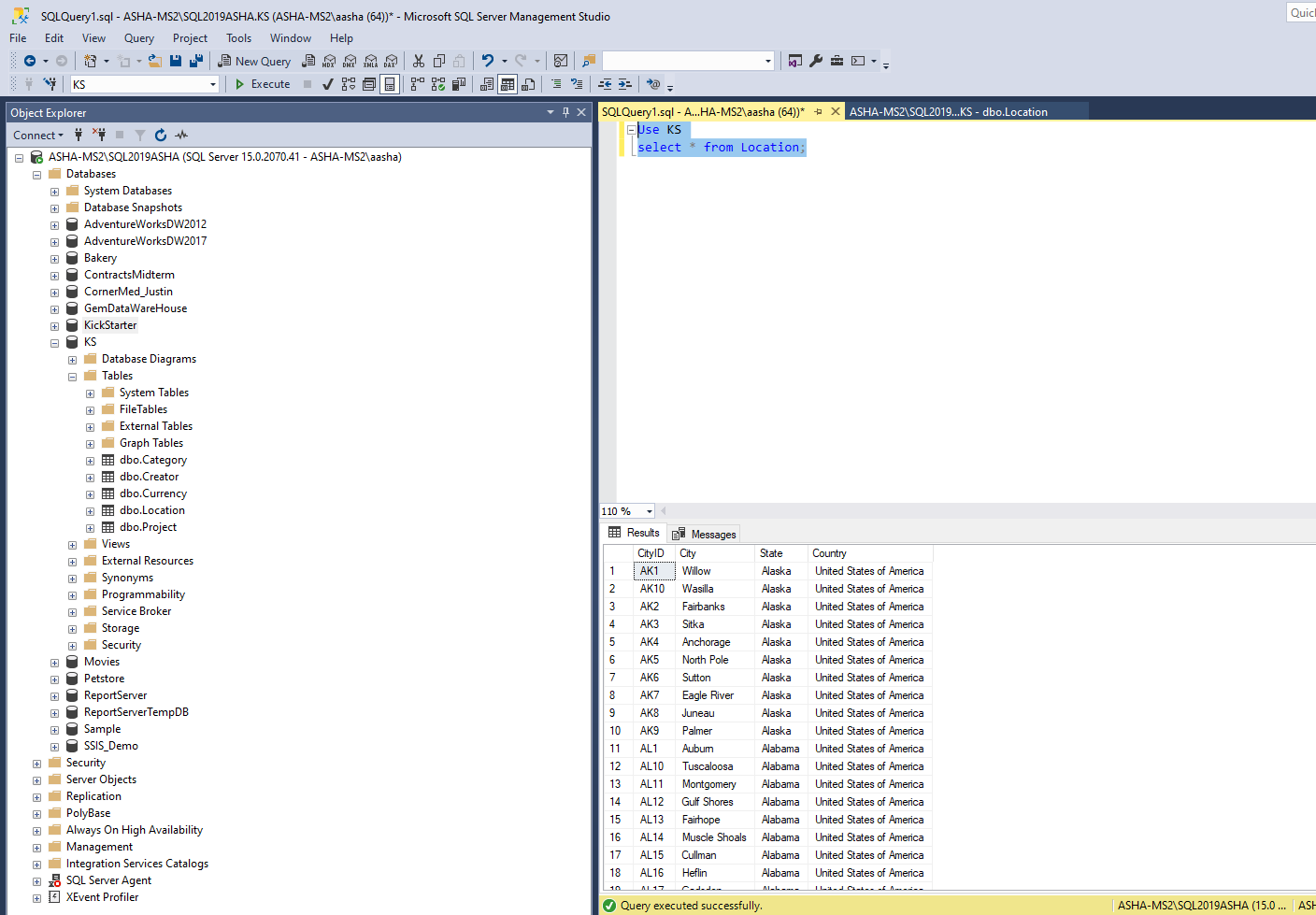
Data Flow – Location Table



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

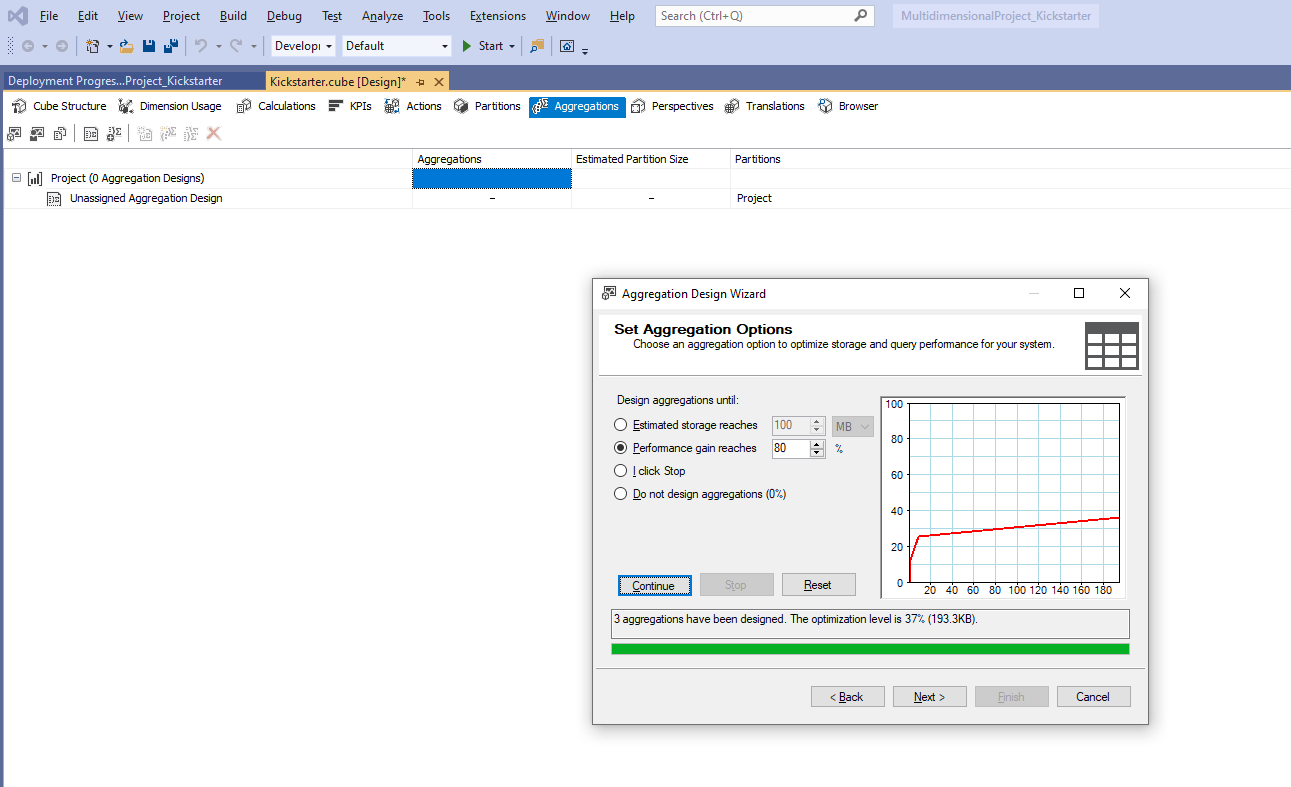


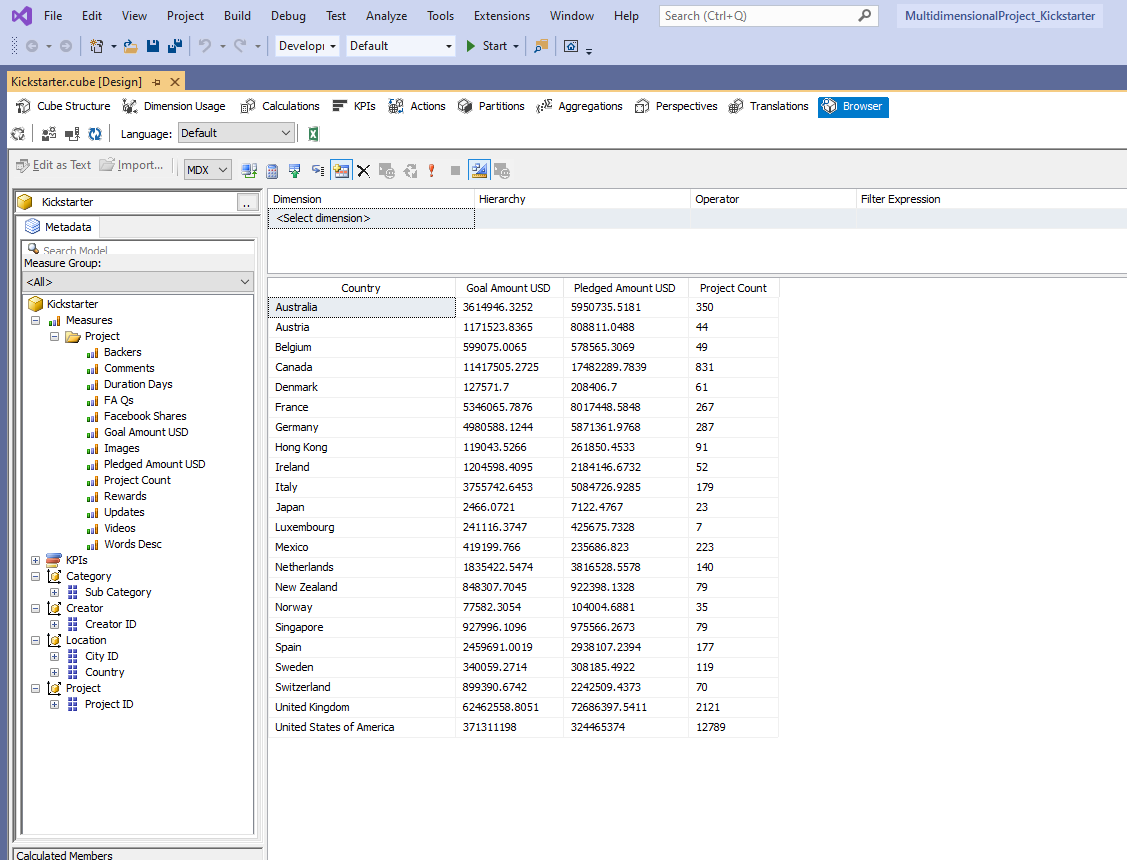


**SSAS-**

**Data Analysis**

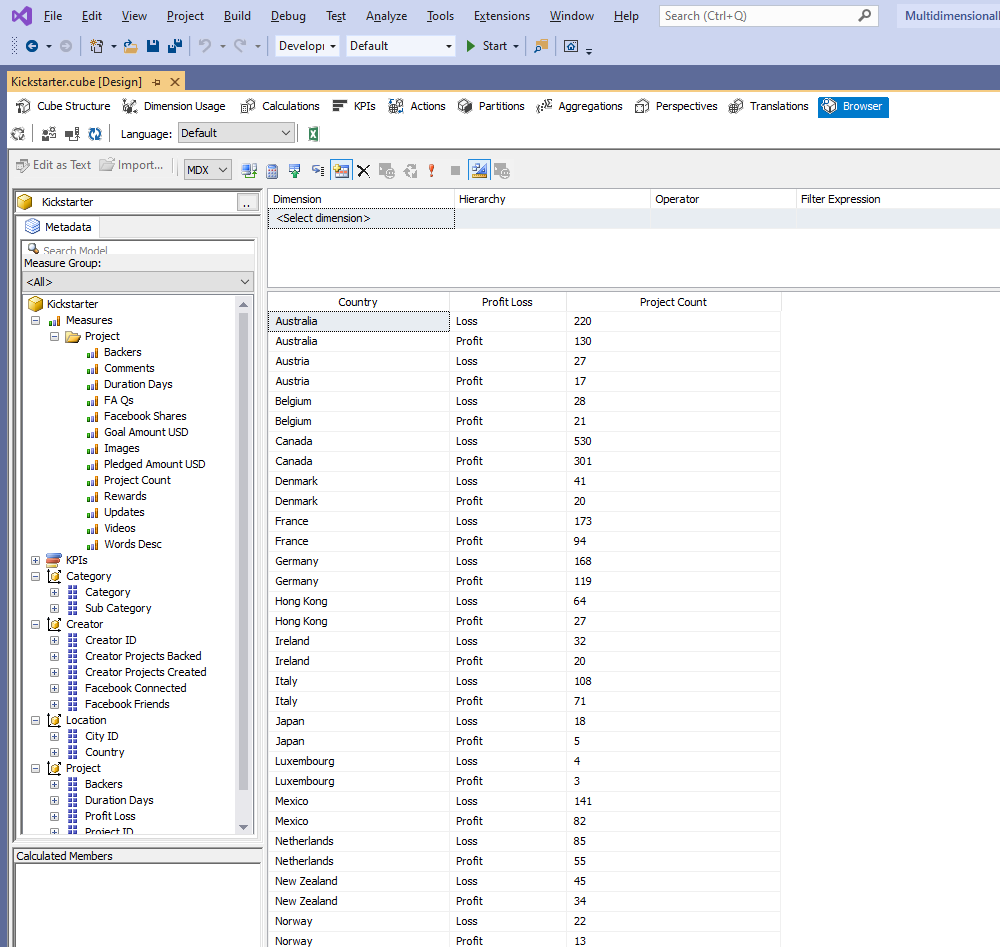
Created cubes using the data exported to SQL to analyze data. Added data source from SQL, created data source views and then created a cube based on the data source view. Processed the cube to generate the final output. Once the cube is processed click on browser tab and add measures and dimensions as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

****

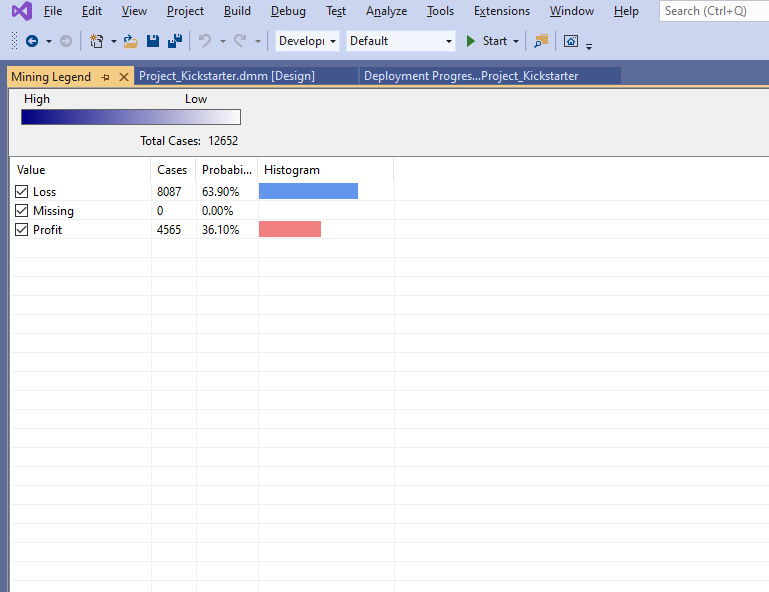
****

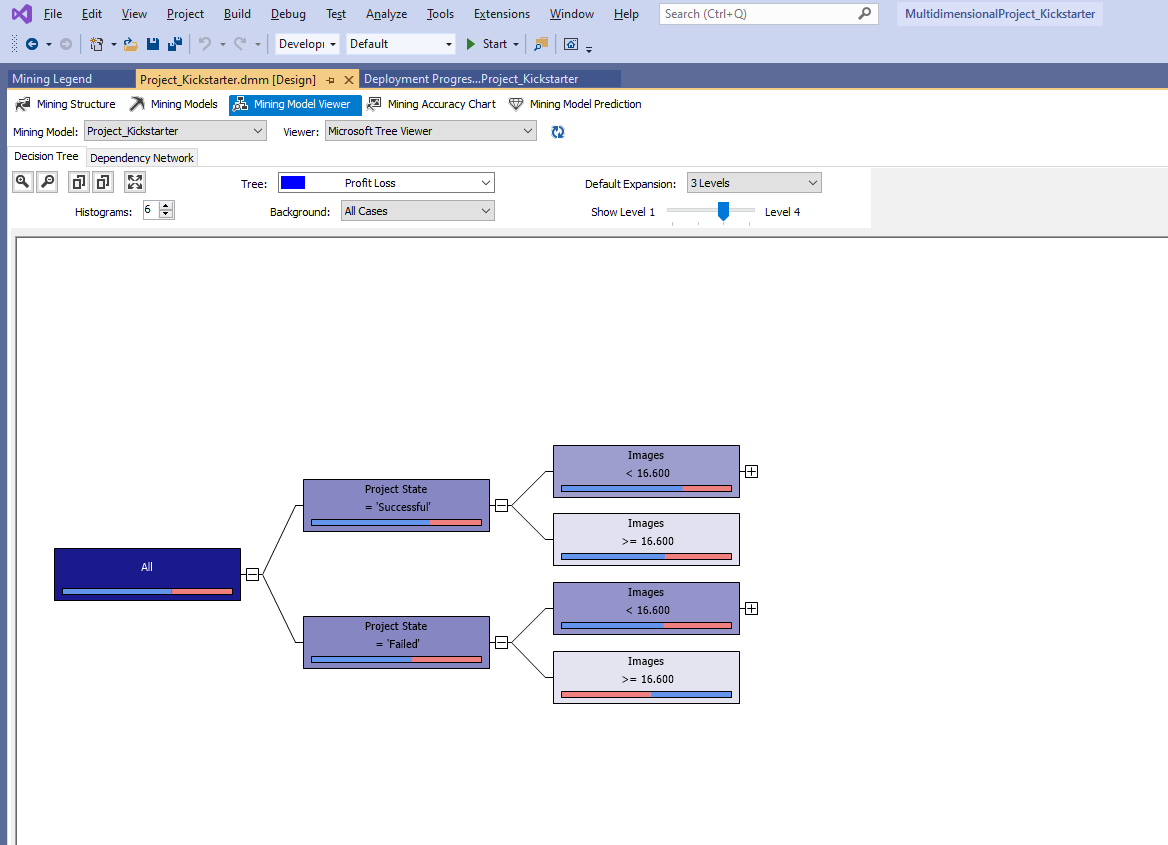
Data Analysis using Cubes – Added new dimensions to cube and re-processed the cube to analyze requirements given. Wrote MDX queries to sort the results from cube. We can add as many dimensions required to get inputs from the data.

****

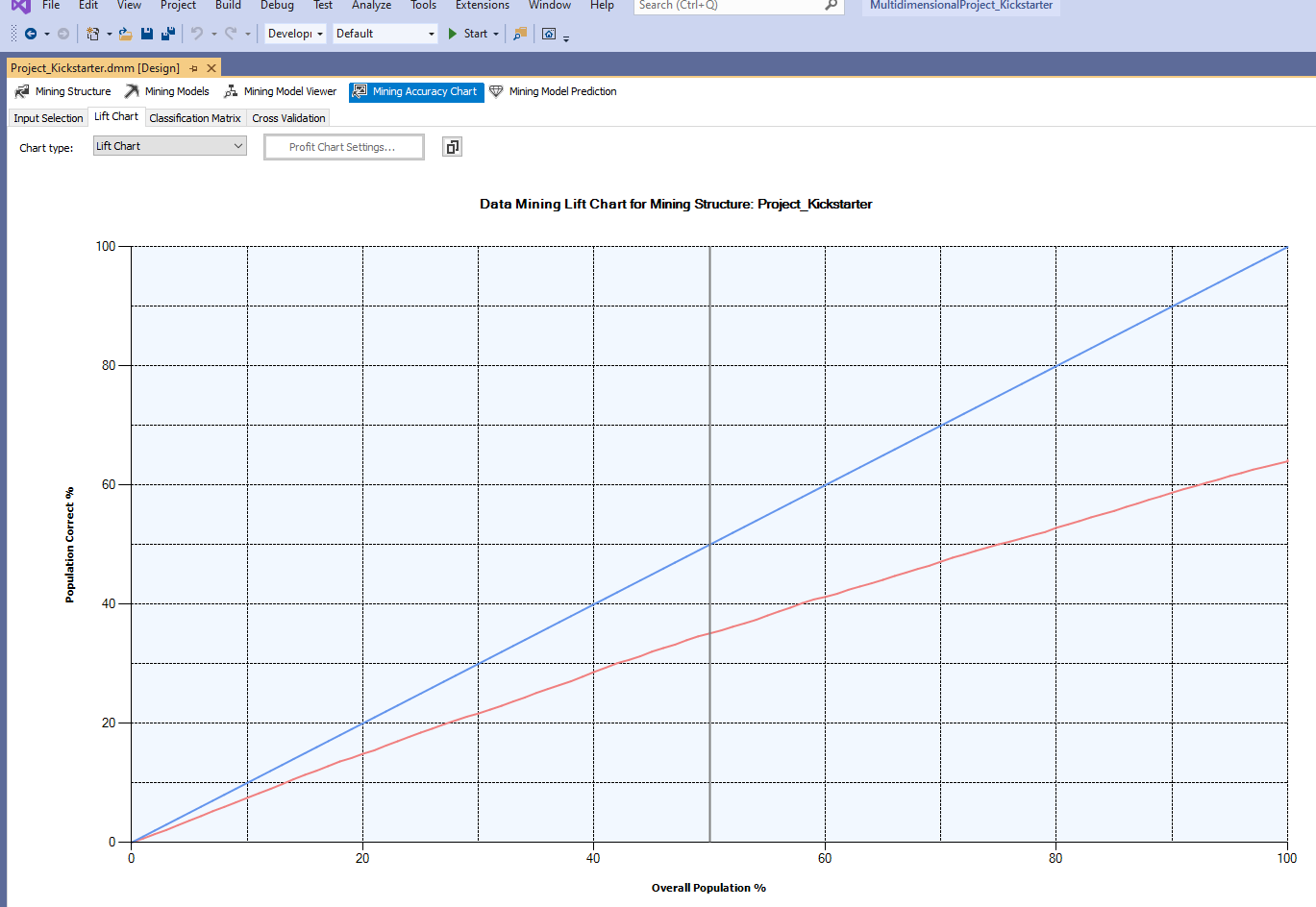
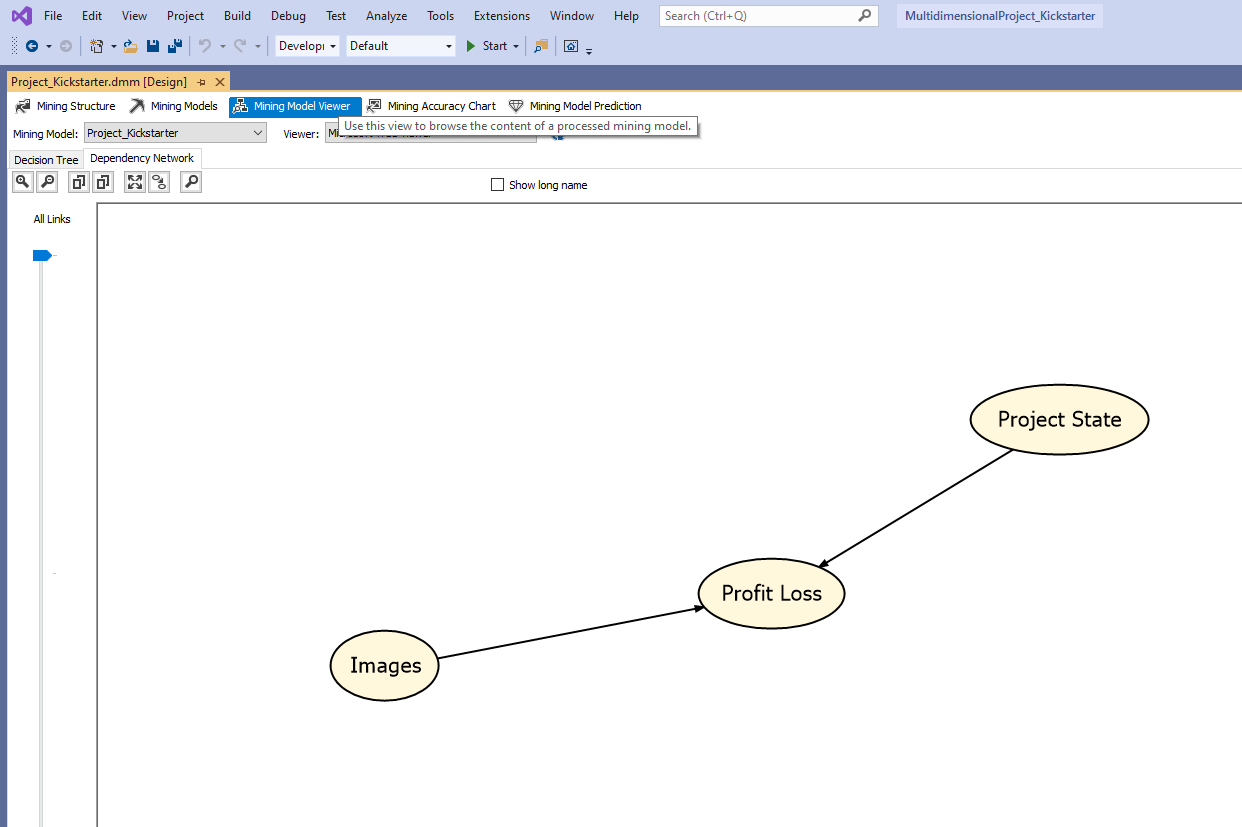
****

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.

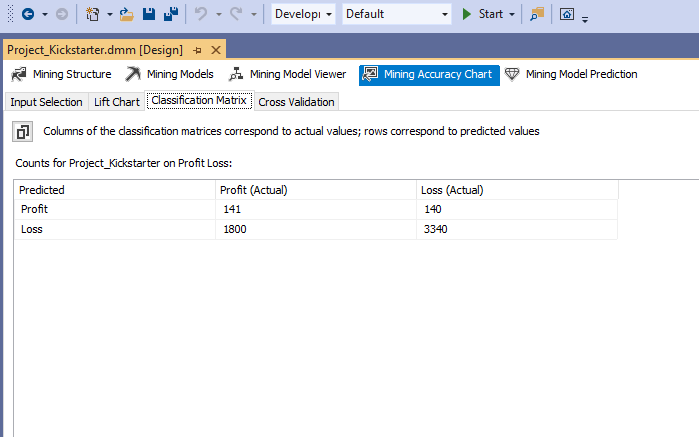
****

****

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.

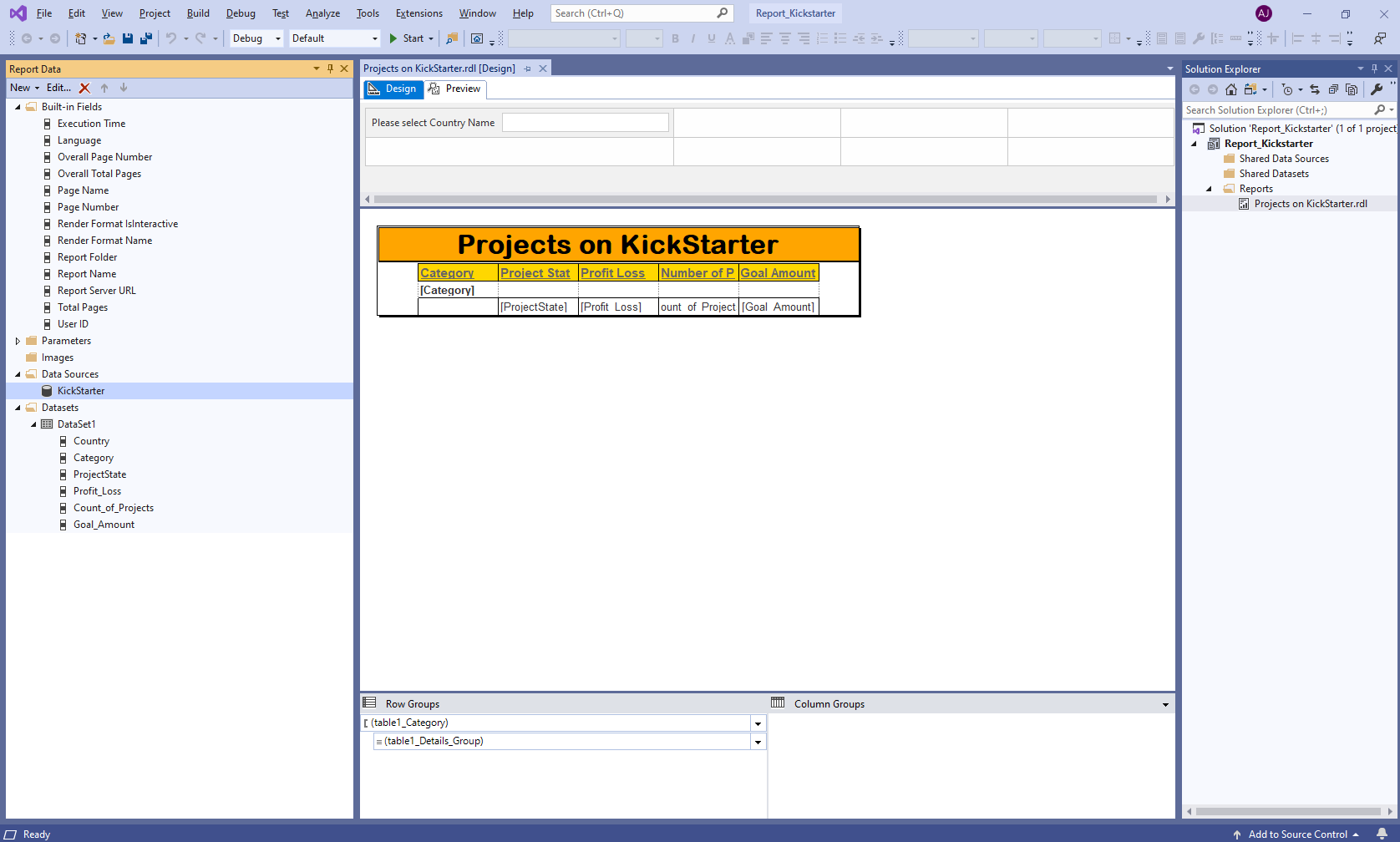
****

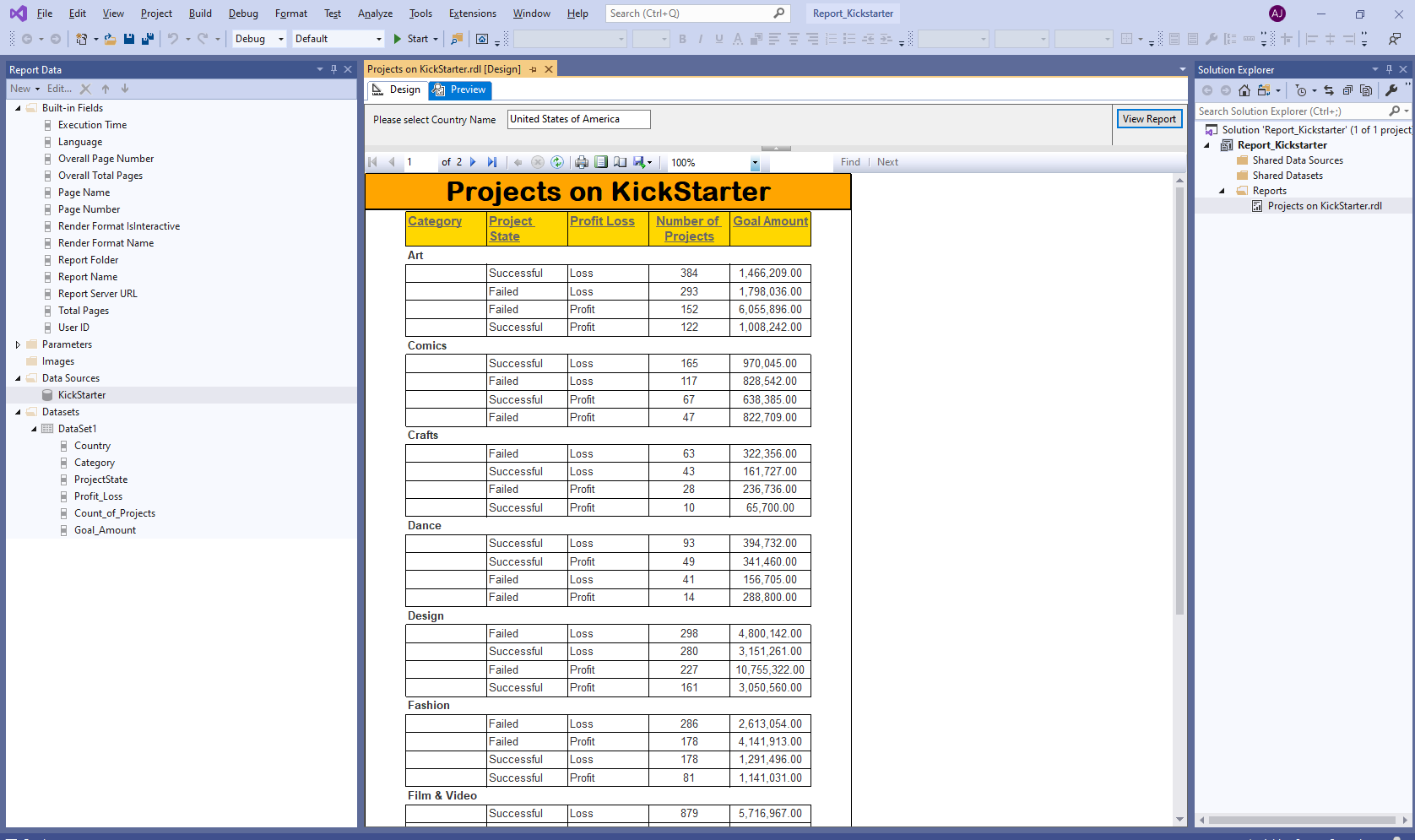
**Analysis from Cubes Created**

* + 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 –**

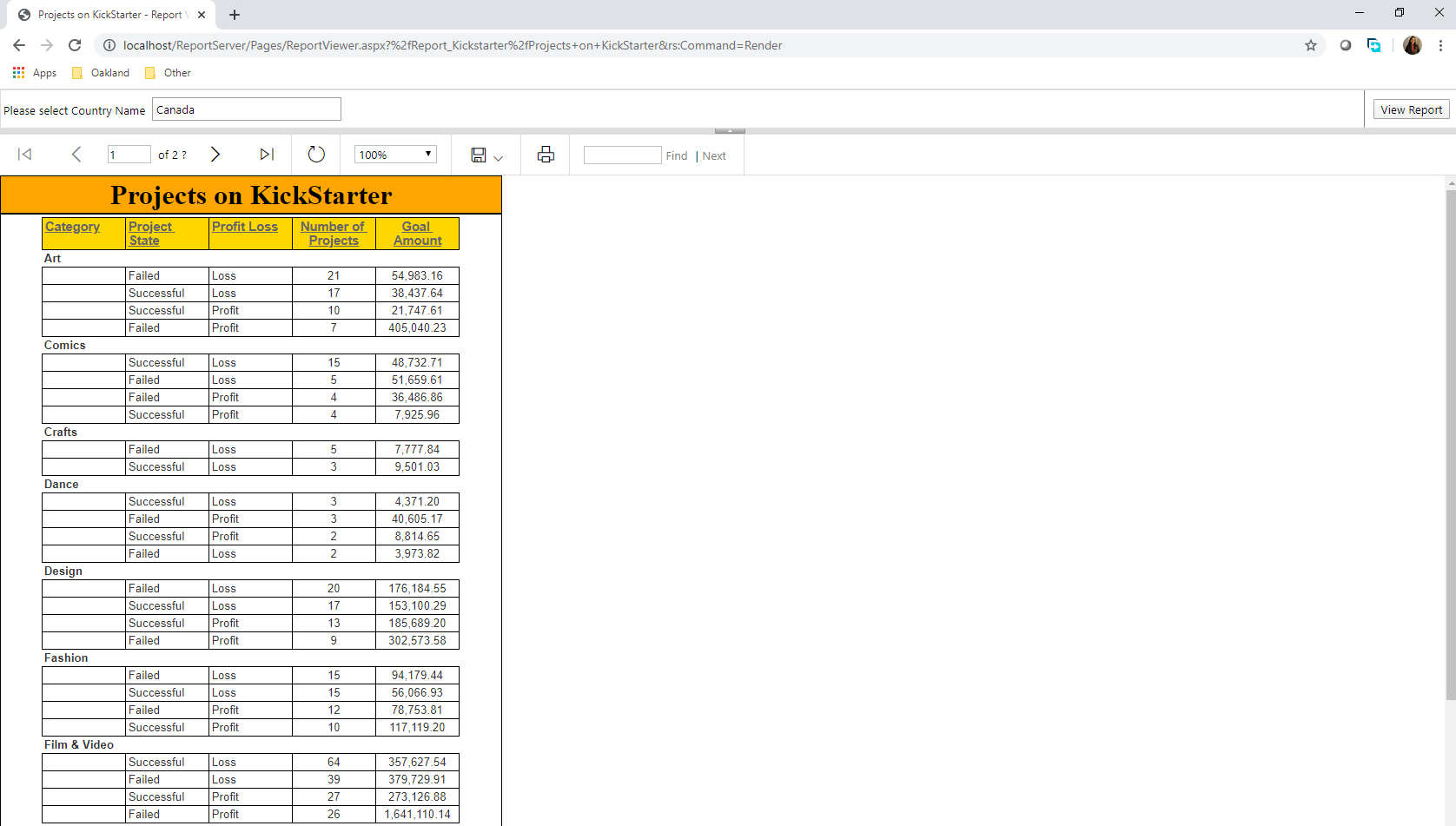
We also created static report using reporting services in Visual Studio which can be uploaded to SharePoint and the team can access data with appropriate access.





Report was then deployed to SharePoint





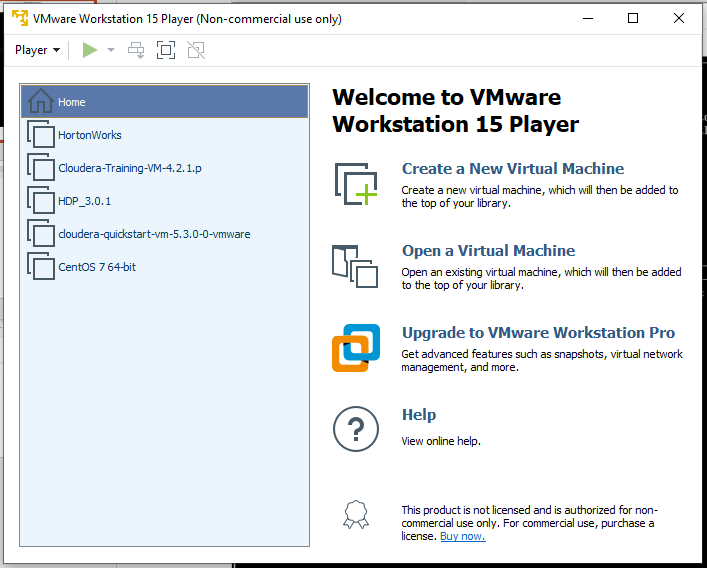
**MODEL 2 – Horton Sandbox and MS Power BI**

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.

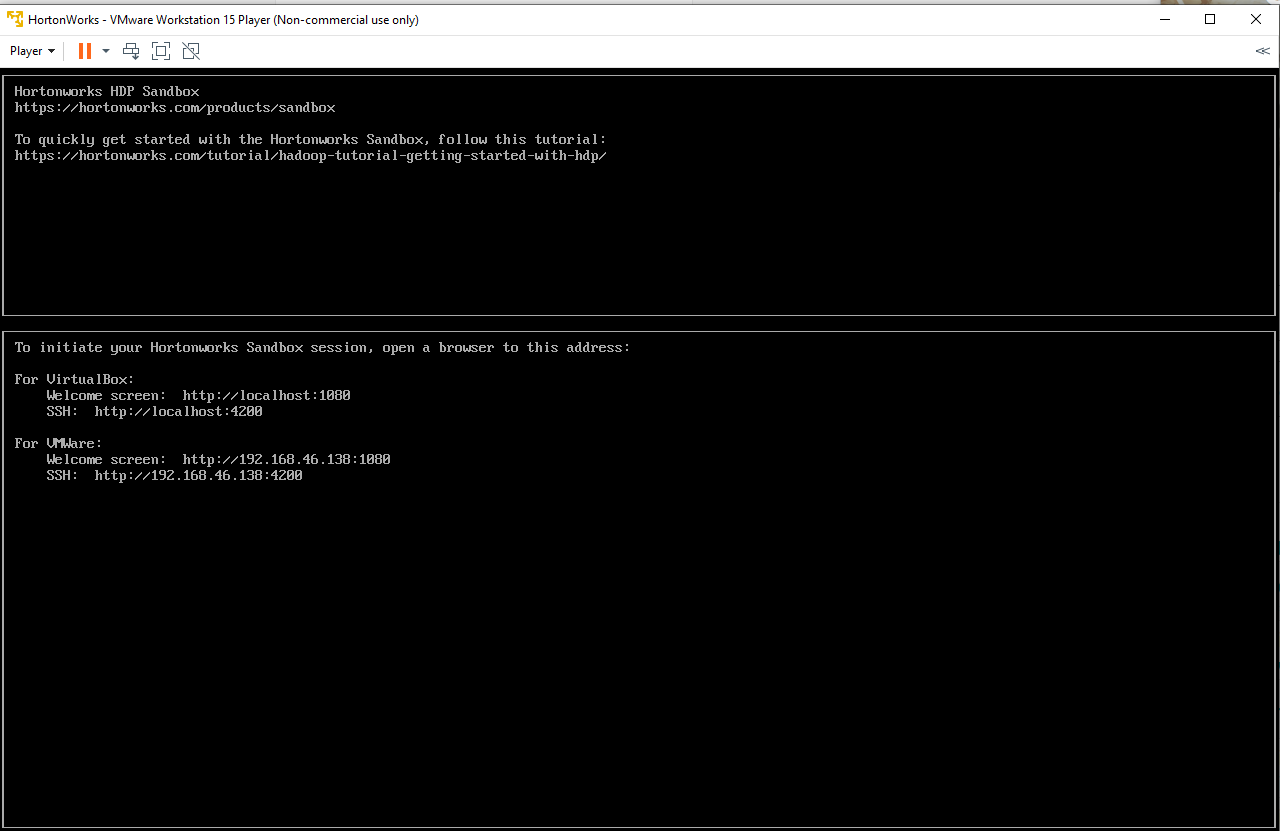
We have connected cityid from project table to location table, to form a many to many relationships and establish relationship. This location table helps business to understand which countries have maximum projects within Kickstarter or gross earnings by each country. The team can also analyze the duration 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**

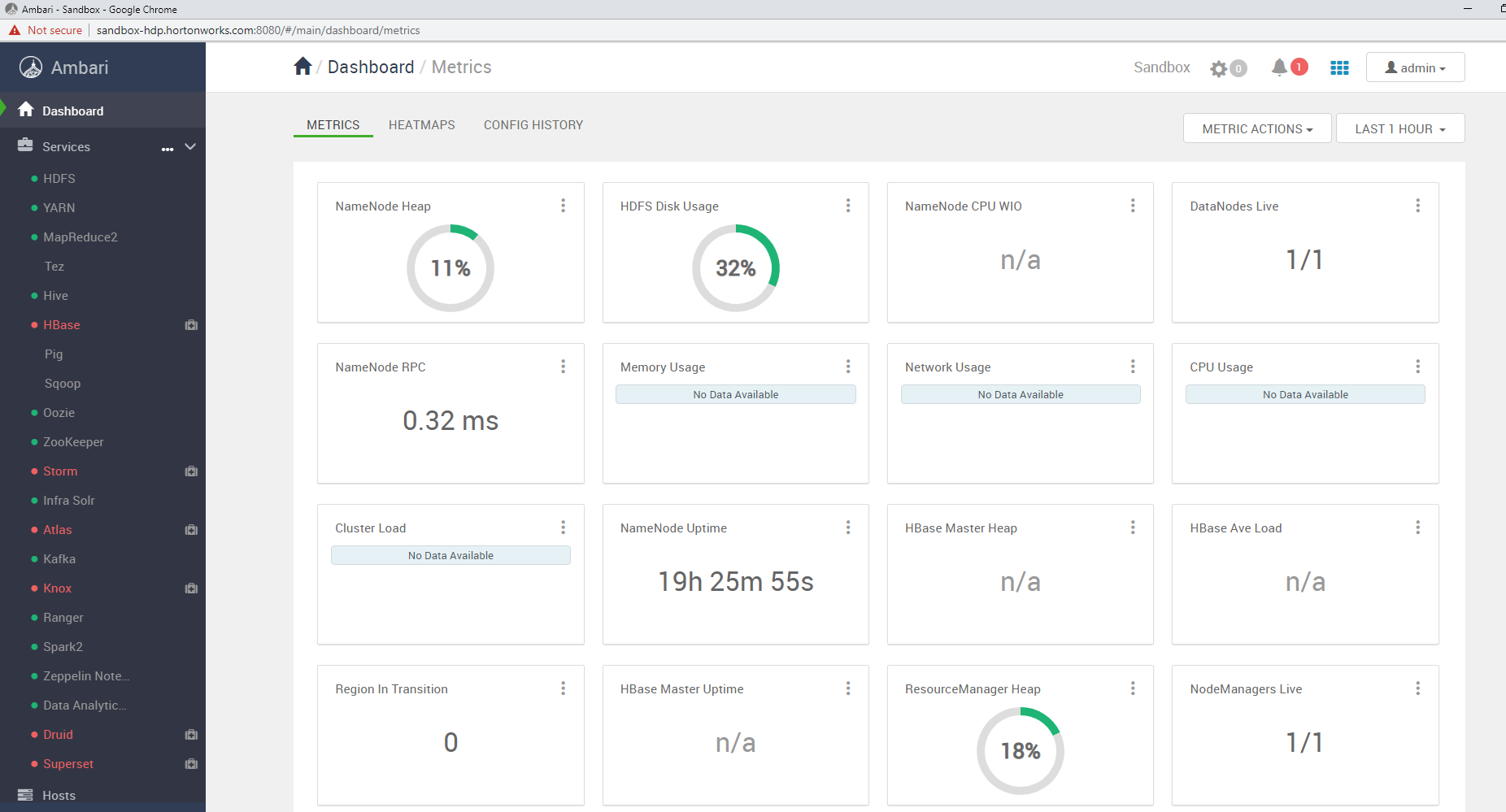
Downloaded and installed VMWare workstation player from the web. After installing player, we downloaded Horton Works Sandbox from Cloudera website. Created a virtual machine in VMware workstation and imported the vdc file downloaded from cloudera.



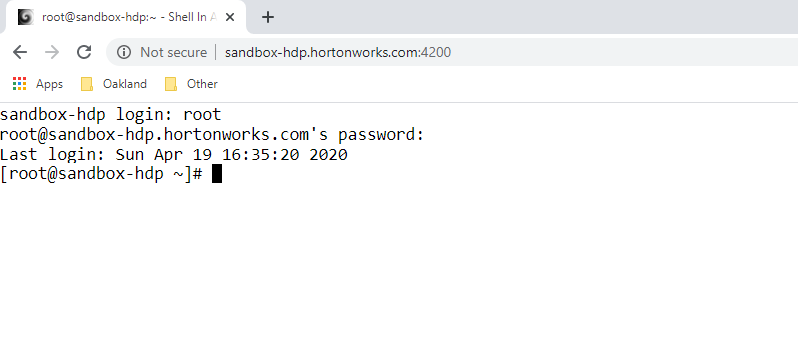
Once the files are imported and the virtual machine fully boots up, we will see a similar screen as below.



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



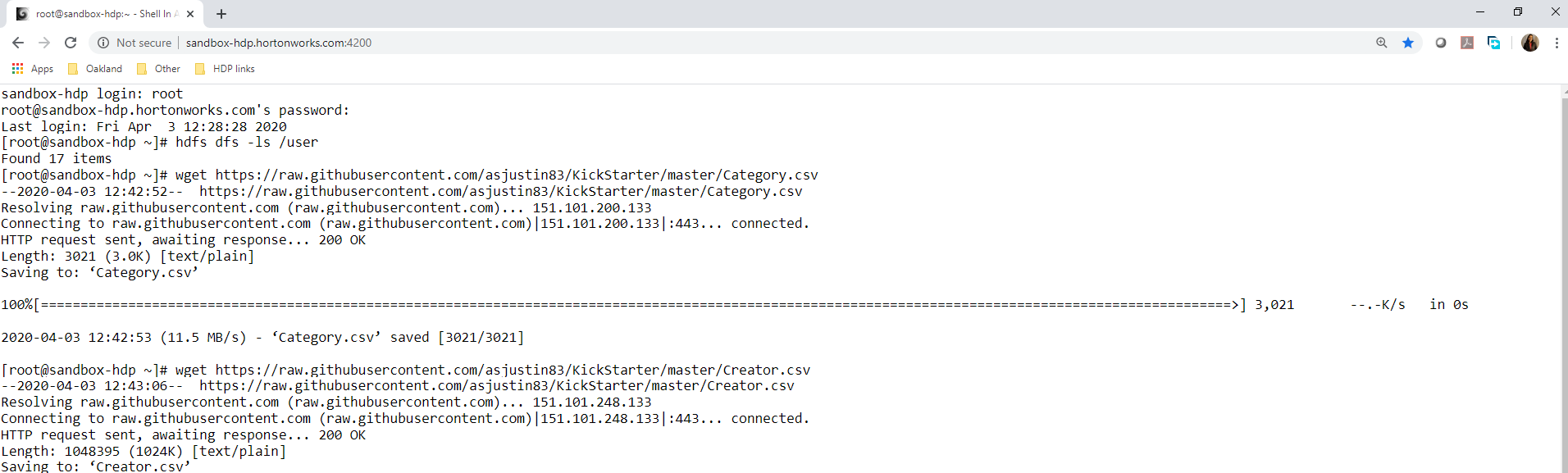
Ambari is the web access view for Horton Works Sandbox. We need to ensure that all the services are up (**YARN, HIVE, TEZ and HDFS**) in running in before we navigate to command prompt. We can add new users from Ambari and provide required access. Command prompt can be accessed at port 4200 of Horton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



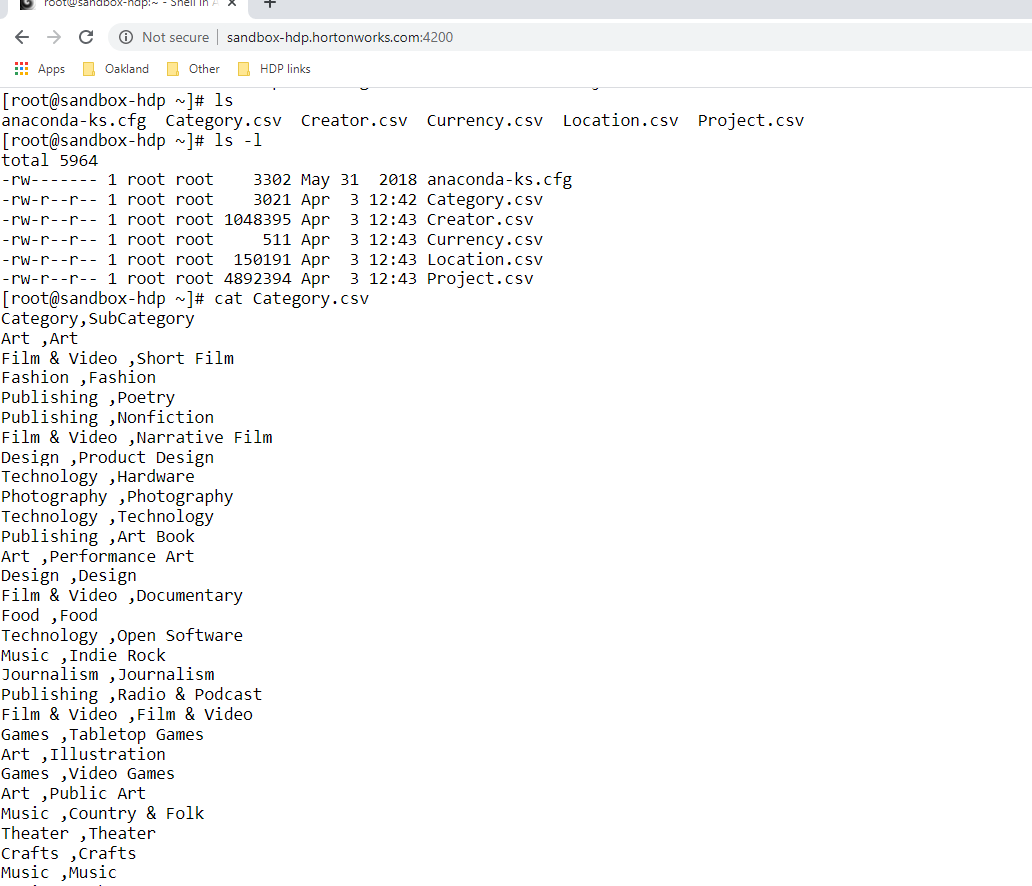
**Data Import to HDFS and Hive - Horton Works Sandbox**

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).

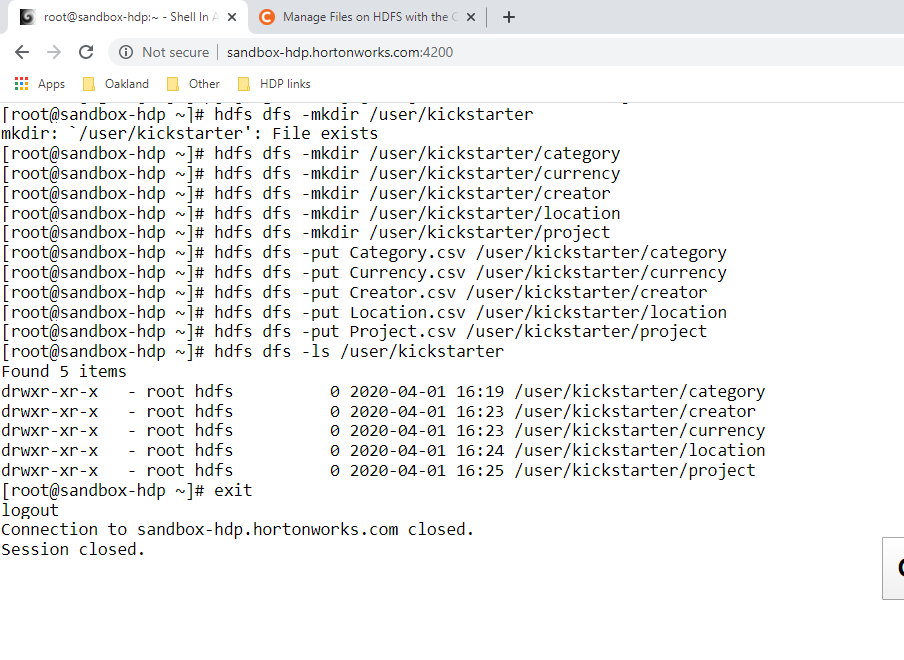
Data from the source file was imported as different .csv files and was saved to a directory in the local drive(/user) of HW Sandbox



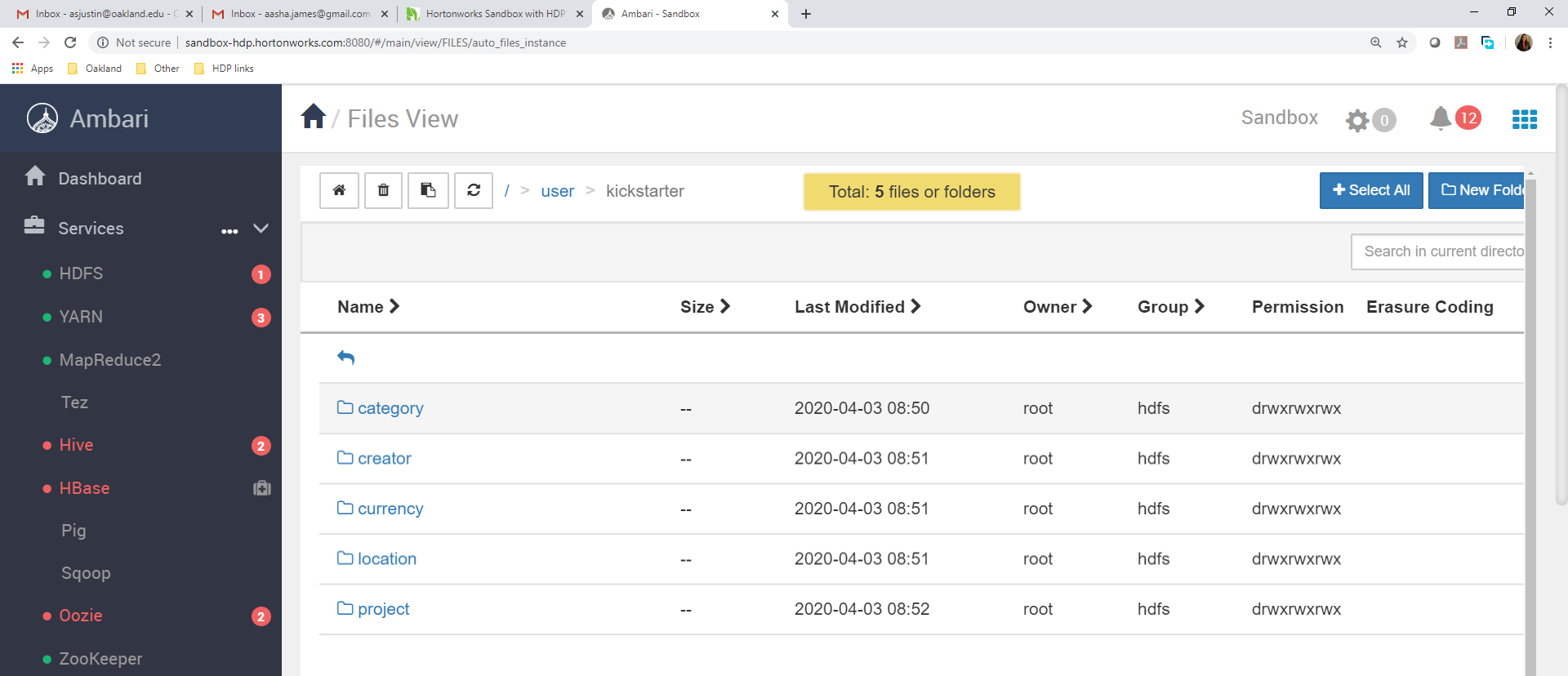
* Checked if the all the files were imported and verified details in the file

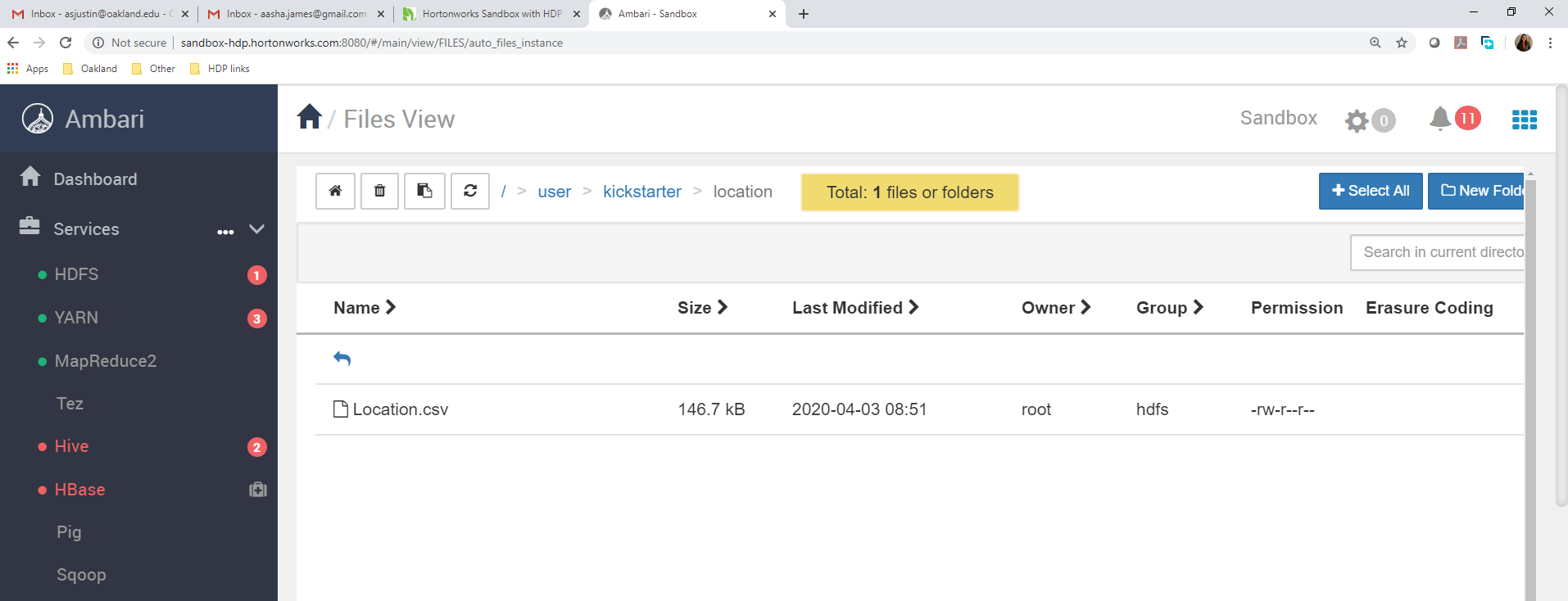


* Created directories in HDFS and moved data from local drive to HDFS directories.

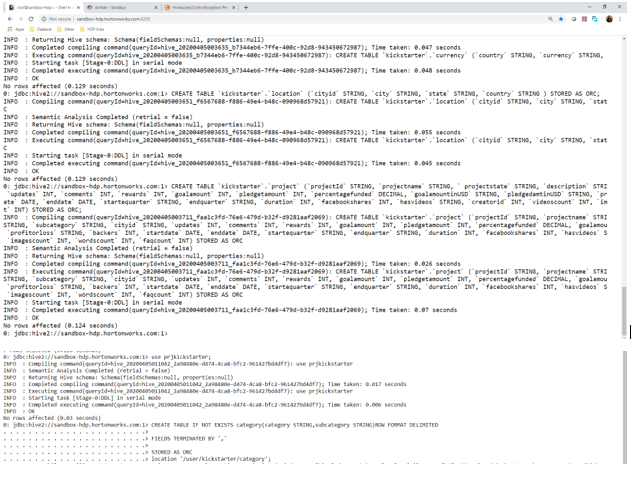


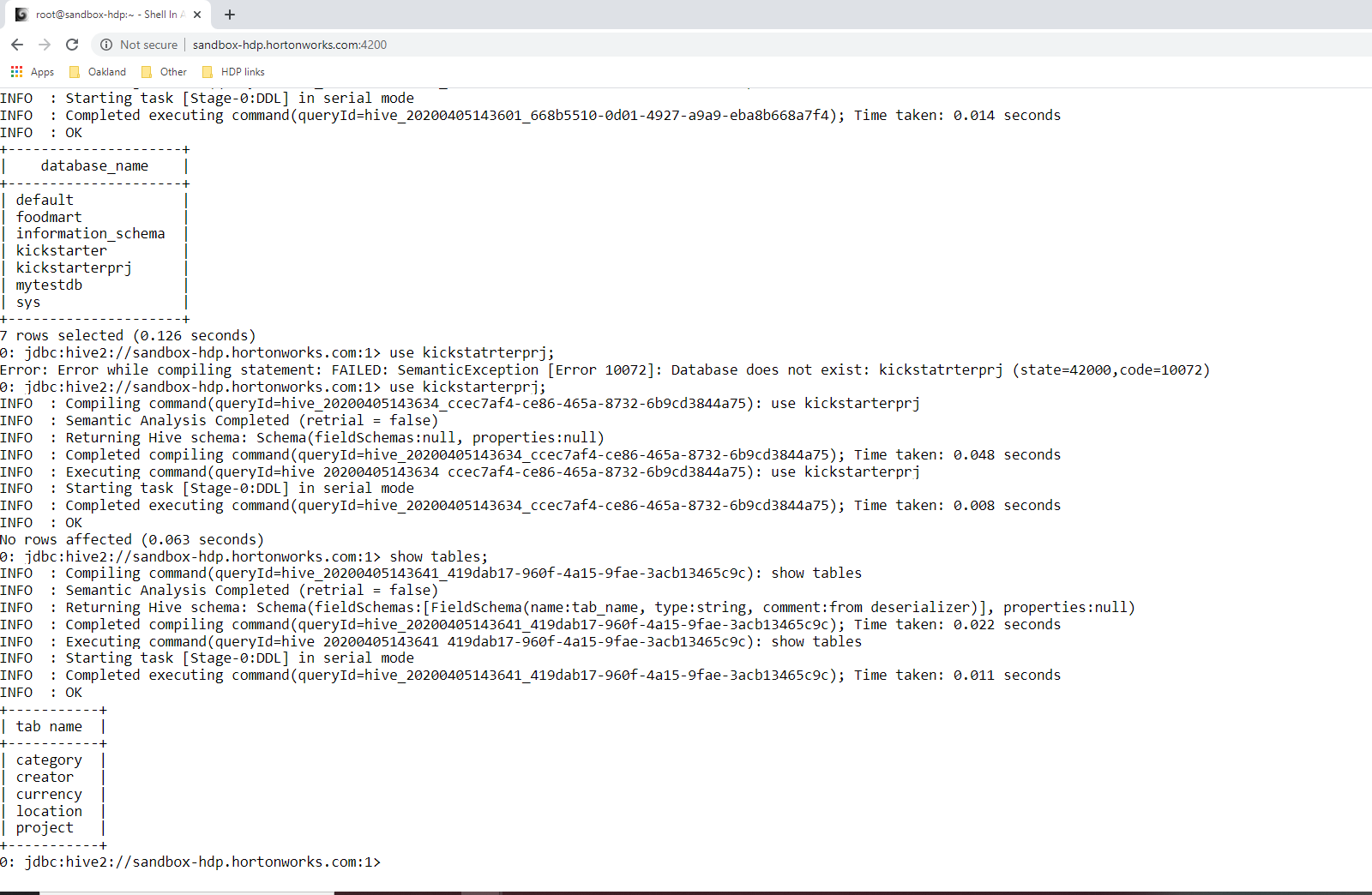
* Verified data uploaded on HDFS using the WebView (Ambari)



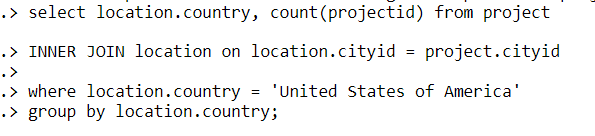


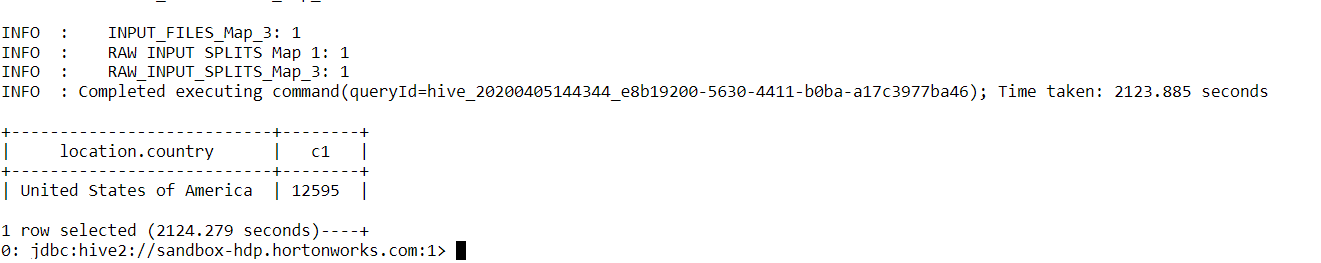
* We can create tables in hive using the web access of Horton Sandbox from Data Studio Analytics by importing the csv files uploaded to HDFS. But for our project we used command prompt to create tables and upload data to these tables from HDFS directories.
  + Created external tables in hive and imported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

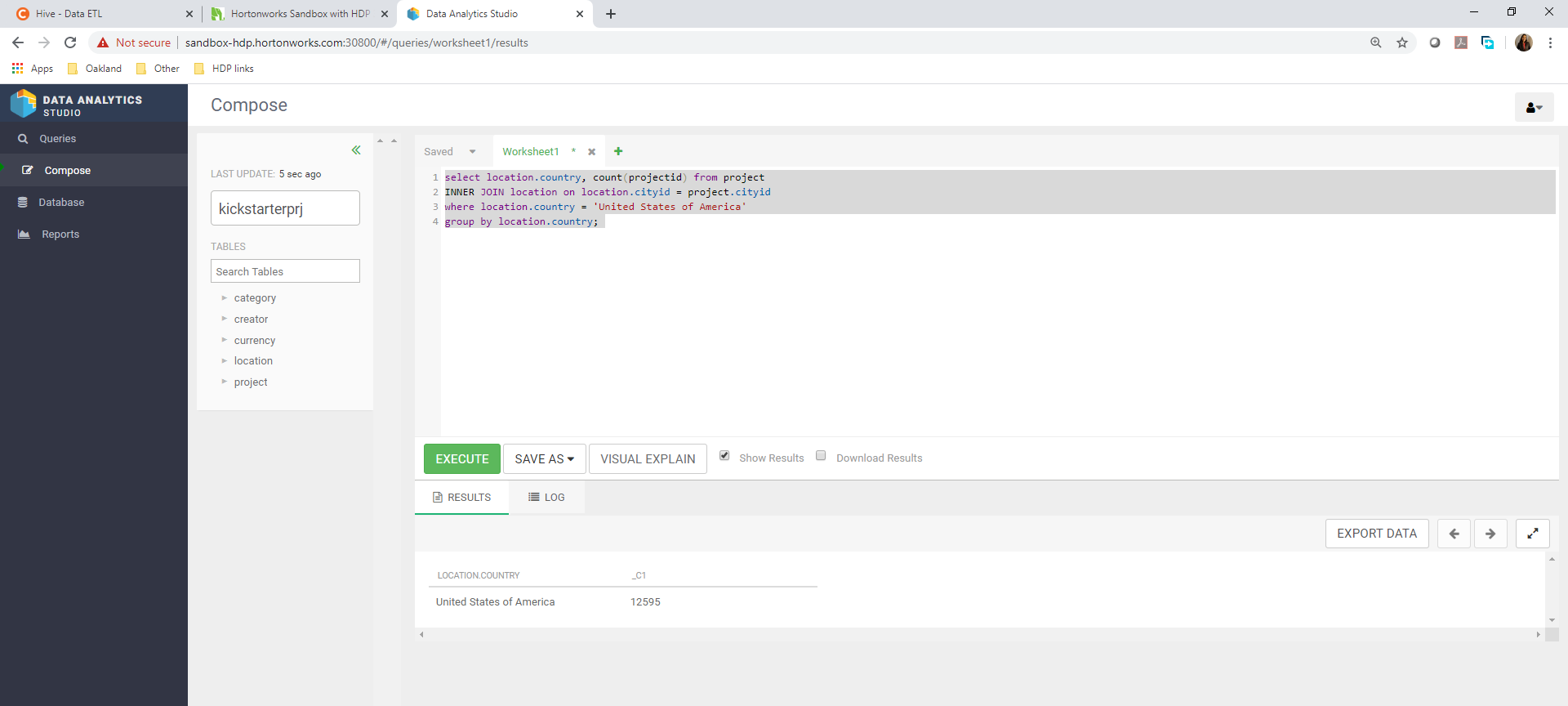




Executed queries both in command prompt and web view to check if the output is identical.







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.

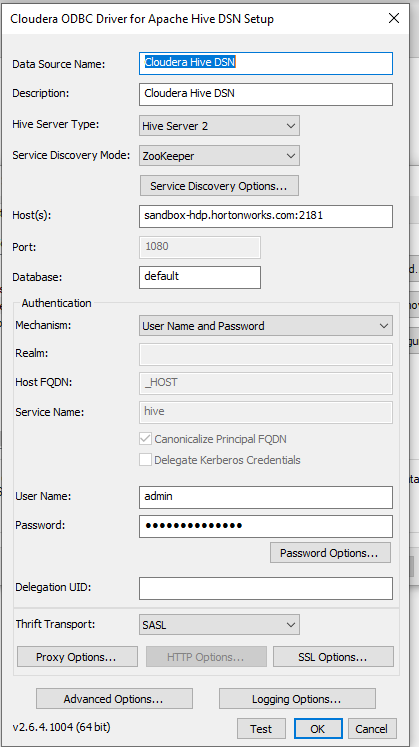
**Data Visualization –**

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.

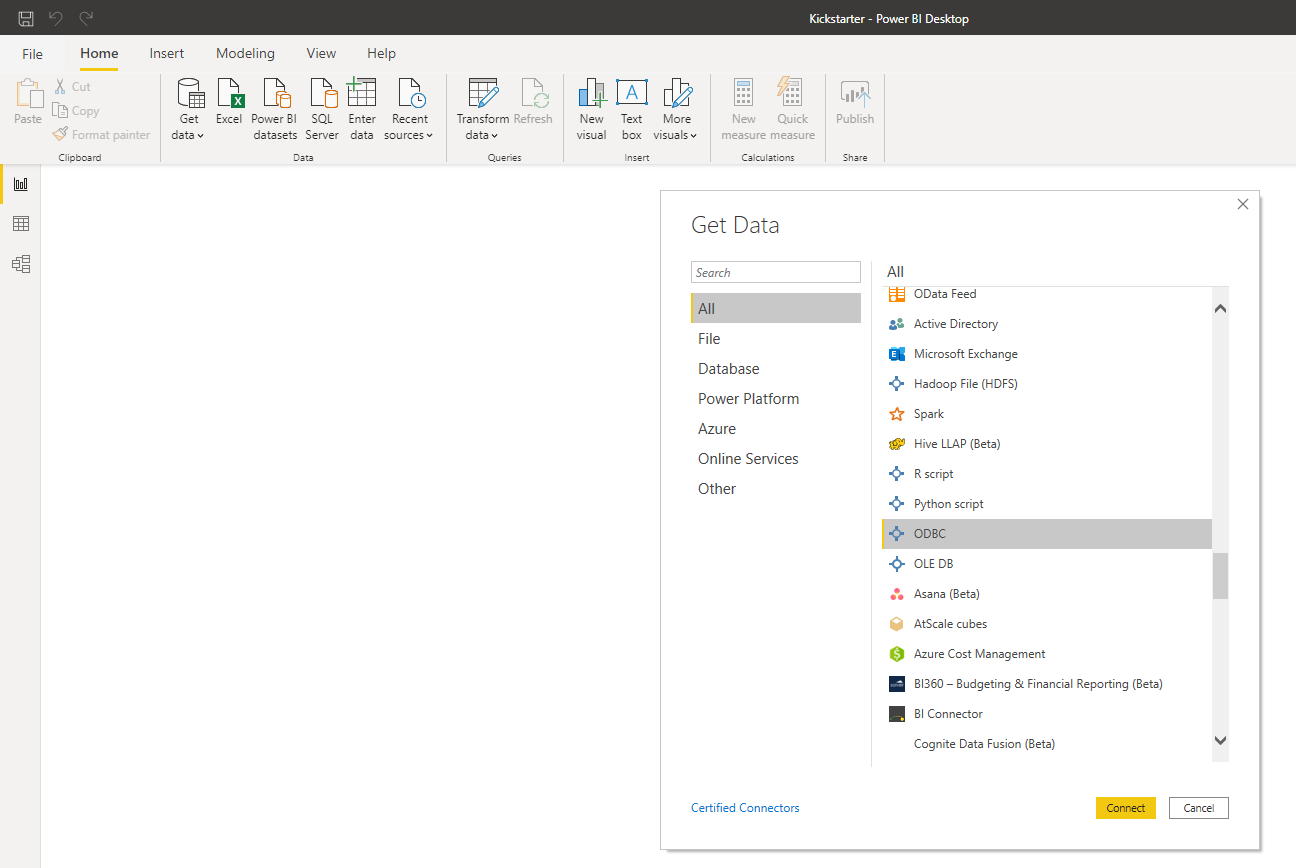
Installed PowerBI desktop version from - https://powerbi.microsoft.com/en-us/downloads. We initialized a connection between Horton Works and PowerBI using ODBC.

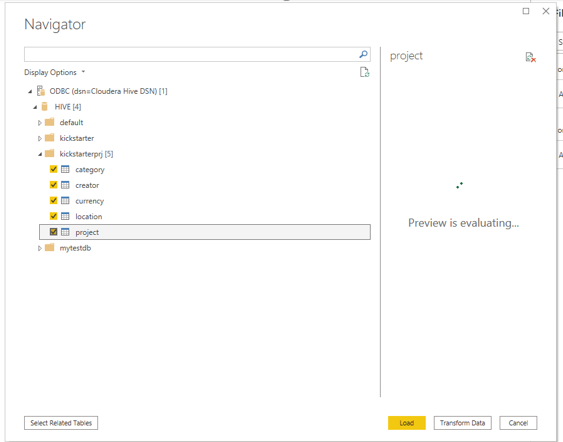
To enable the connection, we need to add ODBC as a connector to PowerBI.

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

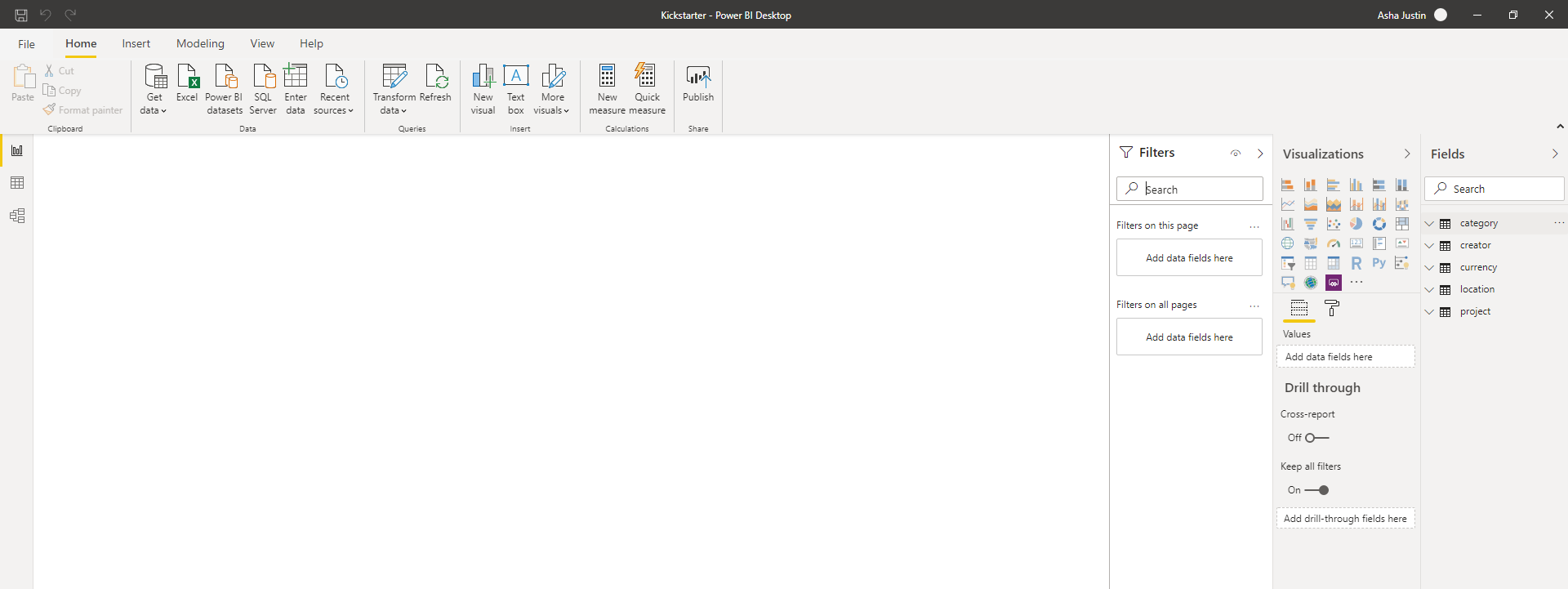


Open PowerBI and click on Get data and select ODBC. Once the connection is enabled, we can view all the tables and databases that are created in Hive.

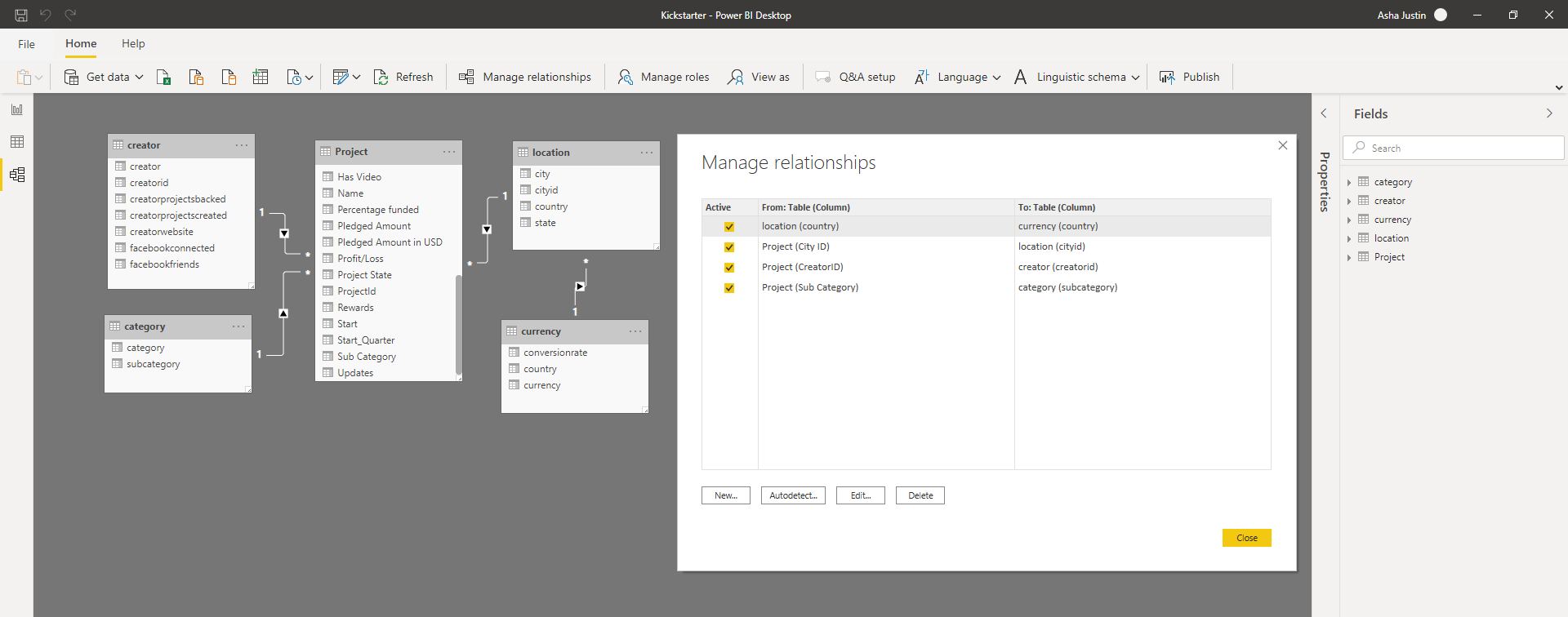




Click on Load and all the tables from Hive will be loaded to PowerBI

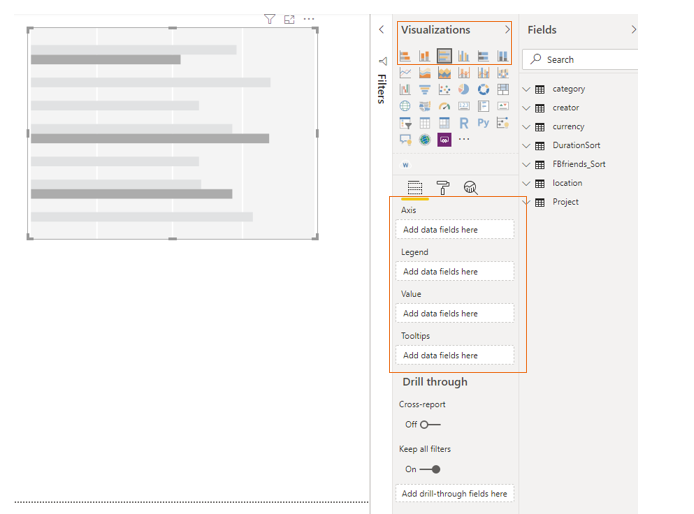


After the tables are uploaded, we need to manage relationship between the tables and modify the primary key and foreign key details (if required) for all the tables in the database.



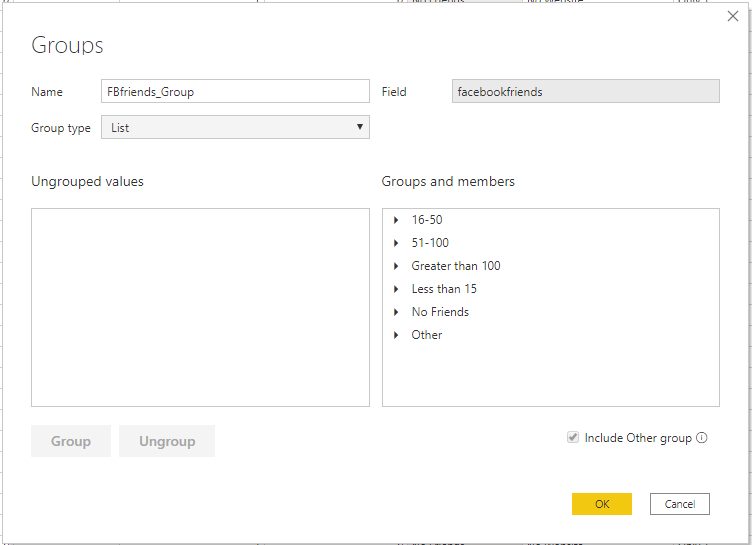
Once the tables are managed as per the relationship, we can now create different charts and analyze management requirements as suggested.

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.

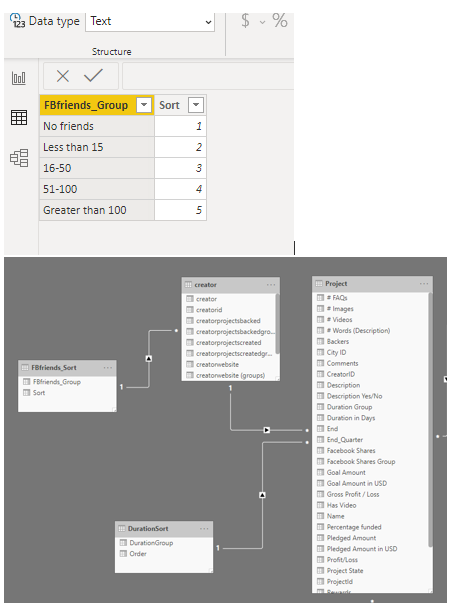


We have selected different visualizations available in Powerbi to create a report which will analyze all the data available in the tables.

We also created groups in PowerBI to analyze numerical data better

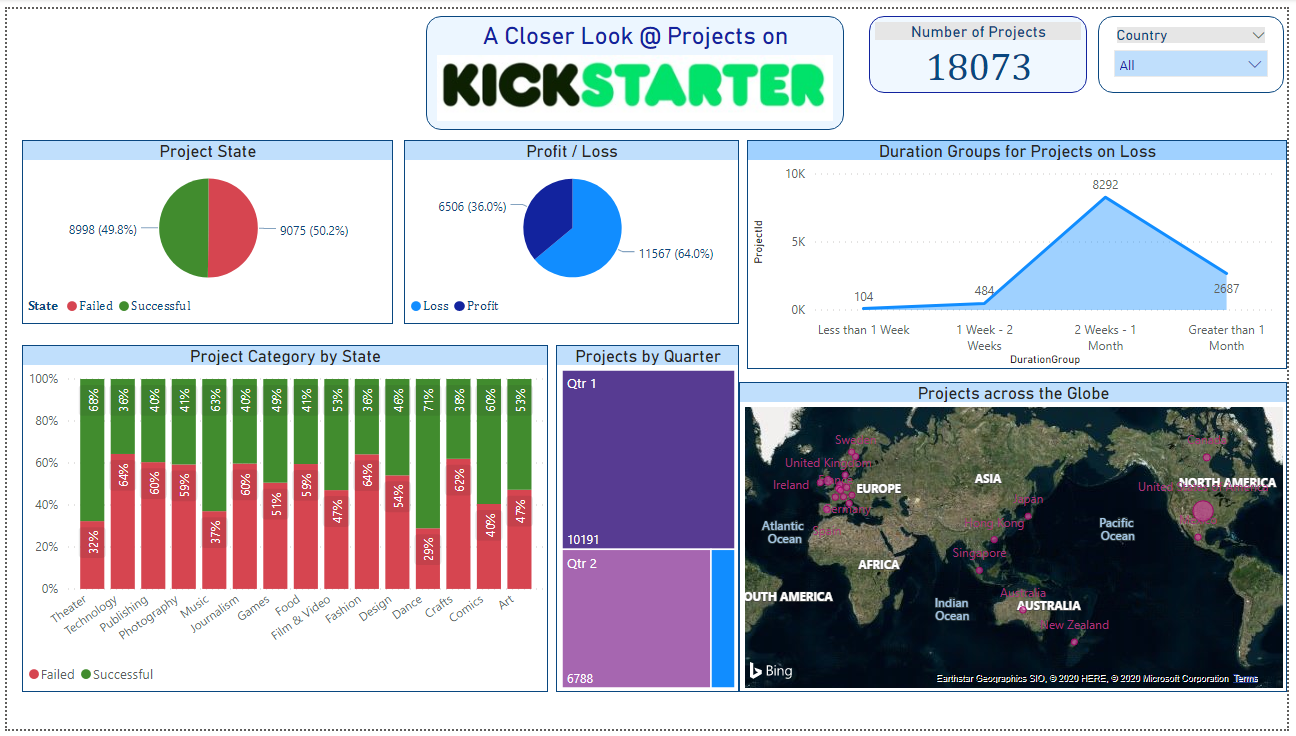


Created separate tables to manage order of the variables which follow an order so that the graphs can updated accordingly. Created relationship with the parent table



**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.

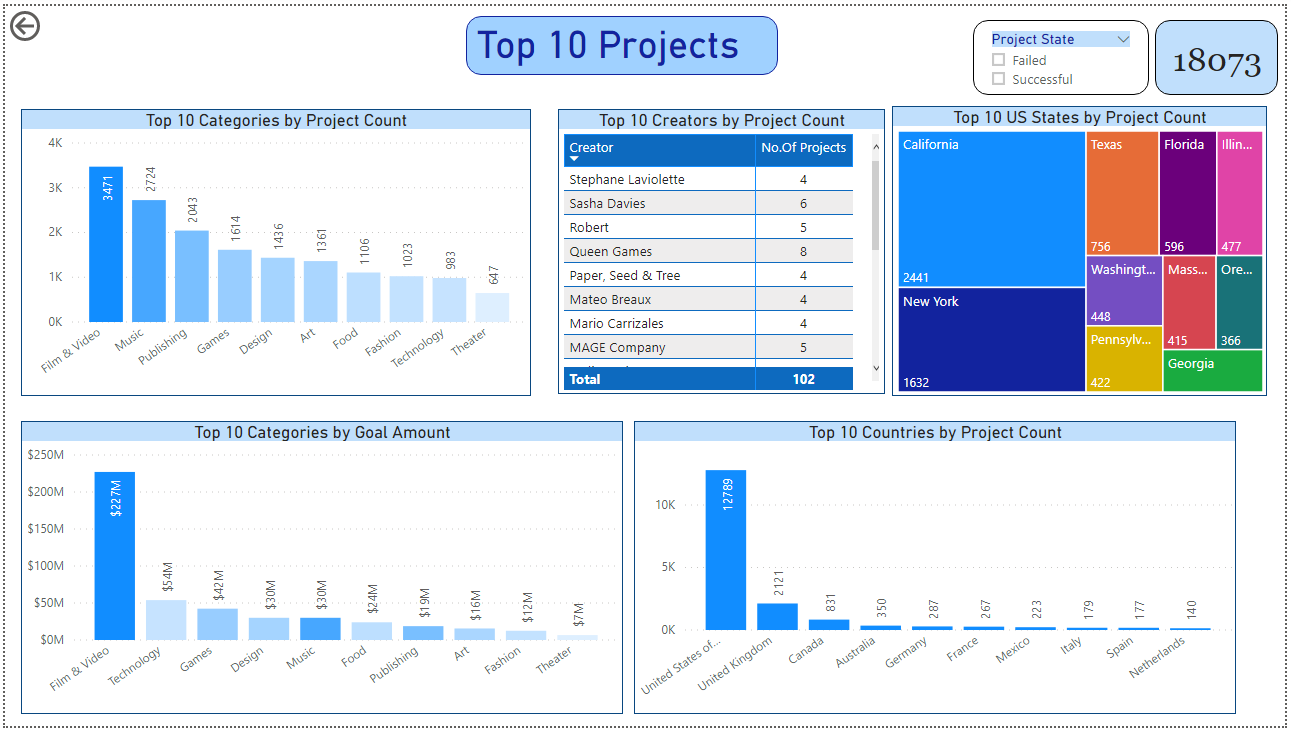
Top successful category by project count – Film & Video (1837)

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**

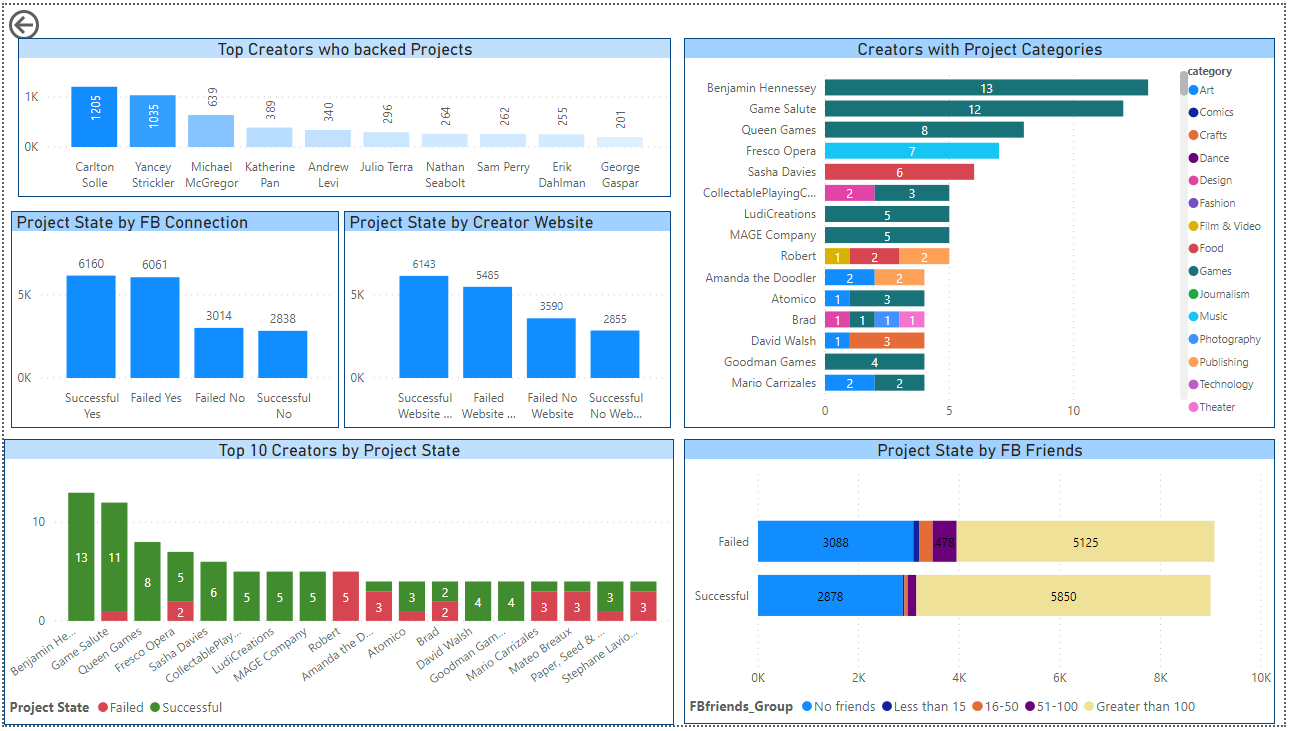
This page will analyze how creators and backers impact the success or failure of the project. We can find who are the creators backing other projects, creators creating projects in different categories.

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 .

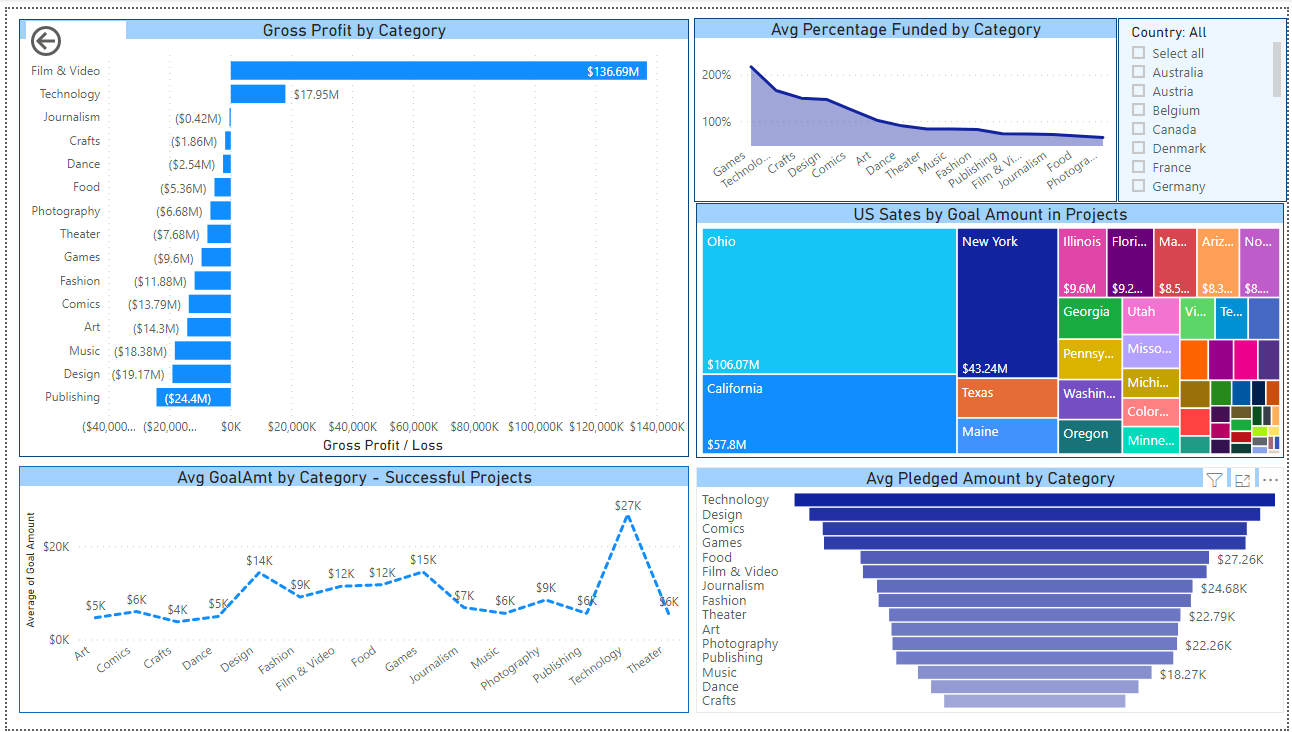


**Revenue Analysis**

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.

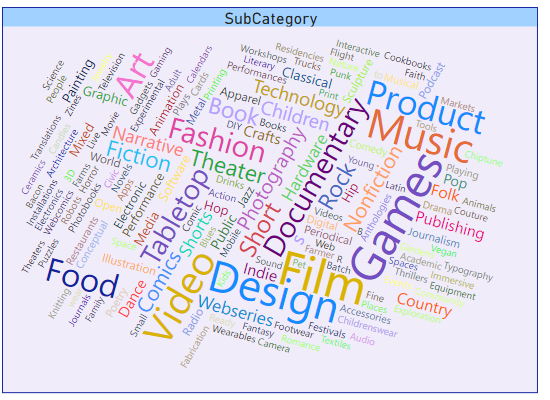
We can also analyze the average goal amount required to create different project types. Technology projects require more money ($27K) whereas art projects are the easiest to be funded.



Major Subcategories

We can import different visualizations from Microsoft App marketplace and can use the same for better visualizations.

Imported WordCloud and added subcategory to analyze the most common subcategories within Kickstarter. Design, Video, Documentary , Games ,Fashion Food and Art are some of the most common subcategories of projects updated on Kickstarter



**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>