Brief on Ad Forecasting Use case:

[Time series](https://en.wikipedia.org/wiki/Time_series) analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

Time series is analyzed to determine the long term trend so as to forecast the future or perform some other form of analysis. There are two key attributes that make a time series unique:

1. It is time dependent. So the basic assumption of a linear regression model that the observations are independent doesn’t hold in this case.
2. Along with an increasing or decreasing trend, most time series have some form of seasonality trends, i.e. variations specific to a particular time frame. For example, if you see the sales of a woolen jacket over time, you will invariably find higher sales in winter seasons.

Time series forecasting is the use of a model to predict future values based on previously observed values.

The key benefits from ad forecasting for the client is:

* Helps in identifying hidden patterns, associations and trends
* The analyses helps the client to customize the ad viewing experience on the basis of the demography and the gender.
* It aids insights into better understanding of the current market movers and gainers and understand which category of ads are not performing as expected.
* The analysis from these can be then used to address problem areas, enhance processes, and improve customer relationships.

Below diagram shows the different steps that occur in Text Clustering:

Collection of Data

Pre-processing

Exploratory Data Analysis

Modeling

Validation of Model Result

Publication of result

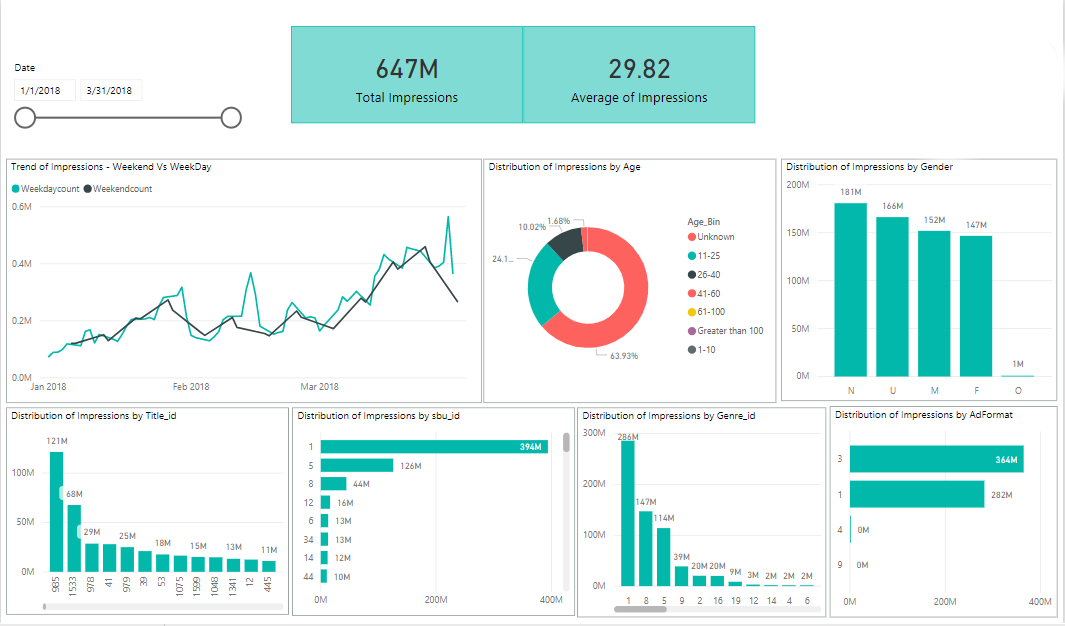
The analytics process for the project is as follows:

Preprocessing:

The data is extracted from the client repository and is checked for initial quality. The number of records that are having no missing values, the entries that are of mixed type in nature and checking for any outliers.

The data that is nearly 2GB in size is incorporated into the local system with a helper function that reduces the loading time of the data from nearly 10mins to less than 3 seconds. The function is generic in nature and is replicable across domains and platforms.

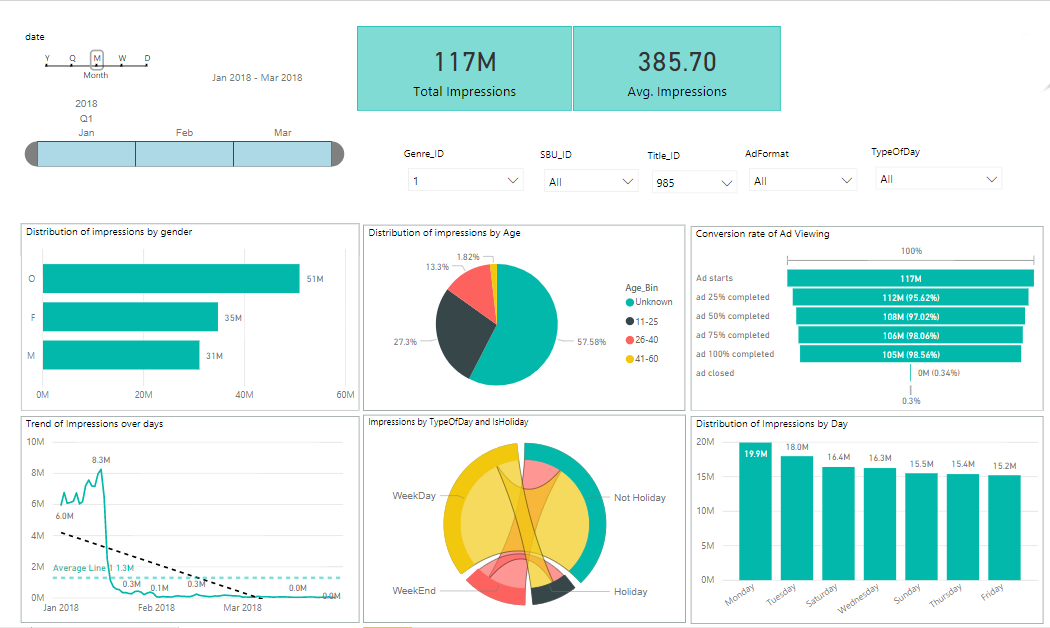
Once the data is loaded, the vital aspect of analytics is to understand how each of the variables interact with each other. This is the exploratory data analysis. The trend of all the major i.e. top 20 ads are taken for the forecasting exercise. Each of the individual trends are analyzed to understand how much variation is seen in the existing data.



Data Summary of the data received, for the period of Jan 2018 till March 2018

Key Insights:

* Overall, there is an increasing trend seen for Impressions from Jan 2018 to Mar 2018.
* Also, the general trend is that there is an increase in the trend of impressions during Month-end compared to the start of the month.
* There is increase in impressions during holidays like Jan 26th, Feb 14th, and March 8th
* More than the weekends, the impressions trend is high during the weekdays.
* Title IDs 985 has the maximum Impressions followed by Title ID 1533, however, the impressions count for Title ID1533 has dropped almost to half of that of Title ID 985.
* Genre ID 1 has the maximum Impressions followed by Genre ID 8. Again, we see the Impressions count dropping to half, between Genre 1 and 8.
* With respect to Age group, we see nearly 64% of data corresponds to users whose age is not known, followed by 24% falling in the age group of 11-25, and 10% falling within the age group of 26-40.
* The impressions trend is the highest for AdFormat 3
* Only for 46% of impressions, we have the correct Gender information.
* However, with the available data the impressions % among the Male and Female, seems almost same.



EDA on cleaned data, for the period of Jan 2018 till March 2018

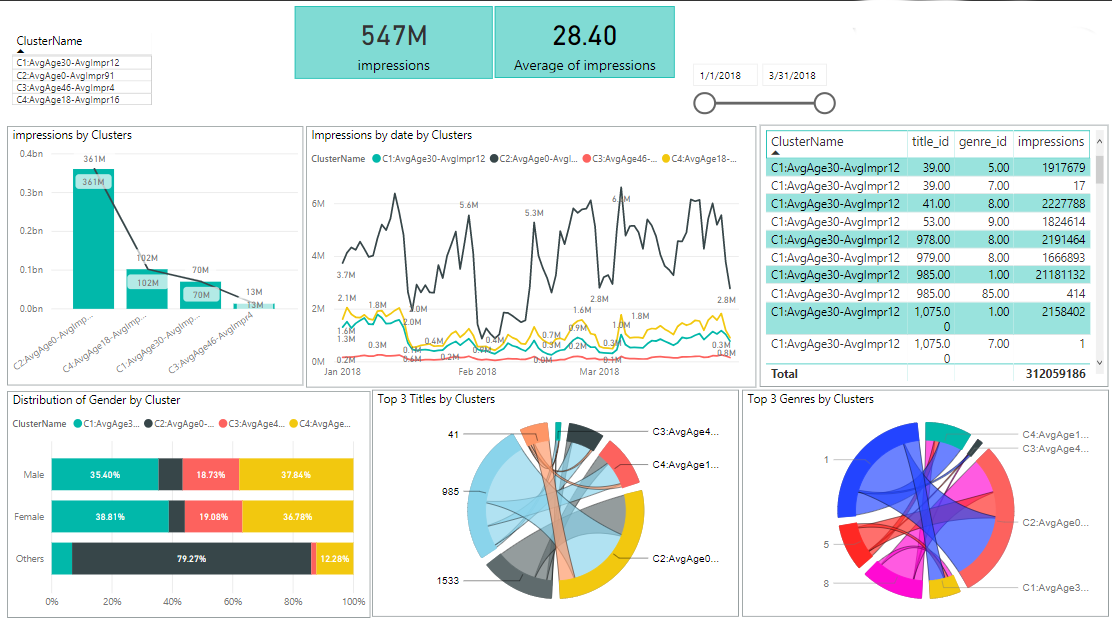
As part of cleansing, other Gender categories like ‘N’, ‘U’ are combined with ‘O’ and Age categories 0-10 and ‘Greater than 100’ are removed from the data

Key Insights:

Filtering on Title\_ID 985 and Genre\_ID 1 which has the highest impressions;

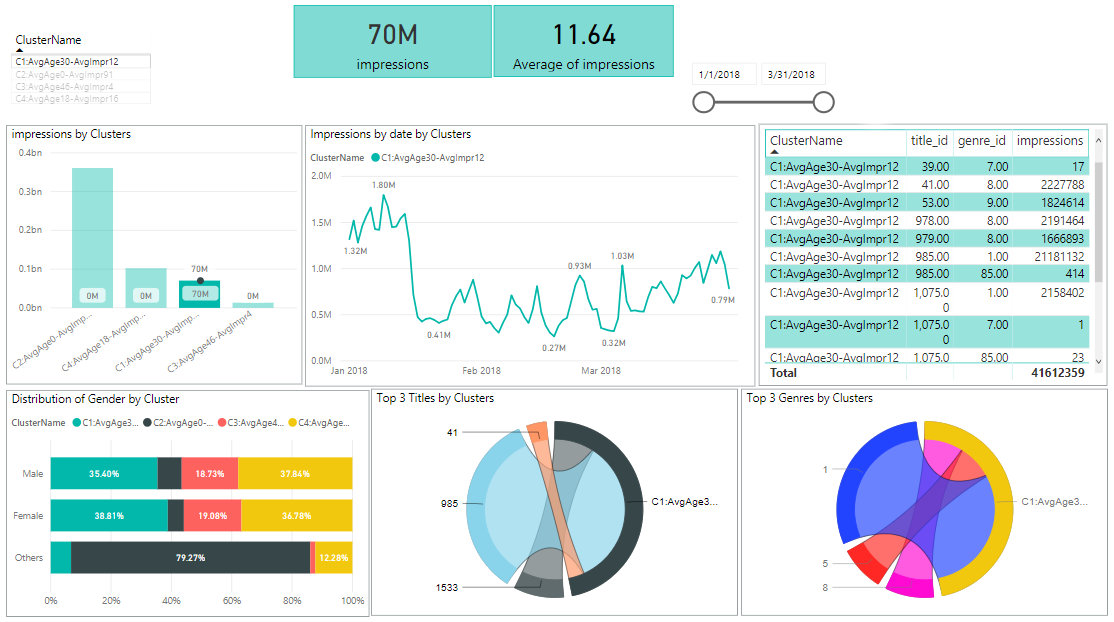
* Male Gender tend to have high impressions, compared to the Female gender. However, this can be confirmed only if we have the unknown gender information.
* Again Age\_bins 11-25 and 26-40 have the highest impressions, excluding the ‘Unknown’ age information.
* In the Ad Viewing – Conversion rate funnel, we see that of 117 M users for whom the Ad started, nearly 105 M of users have continued to watch the Ad till its completion. Also, we see an increasing trend in the % of users who have stayed to watch it from 25% to 50% to 75% till the 100% completion. And the % of users who have manually closed the Ad is very negligible (0.3%)
* However, for the Title\_ID 985, under the Genre\_ID 1, the impressions were high from Jan 1st till 14th 2018, after which there is a steep drop in the impressions count.
* Also, we see that the impressions have been high on Mondays and weekdays showing high impressions trend than weekends.

The K-Means algorithm was used to identify clusters of users who frequented on a particular ad type and shared common tastes. This resulted in four clusters.



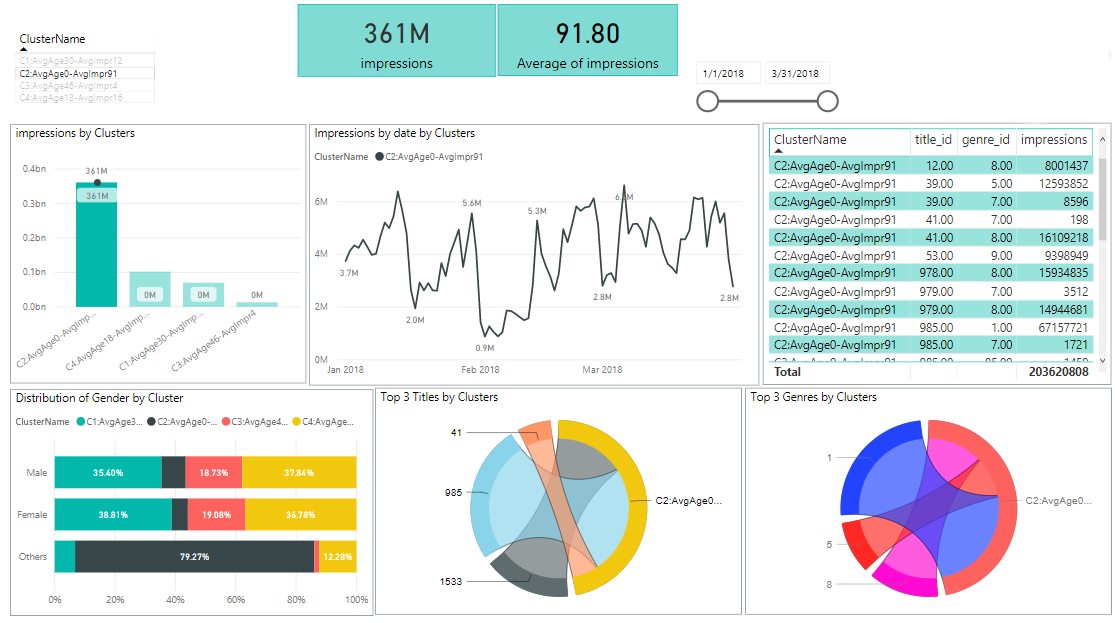
For the cluster C1:

* We see the average of impressions to be ~12 impressions/day.
* The top 3 titles are 985, 1533, and 41 which falls in the genres 1, 8 and 5.
* We see steep drops in Impressions, in this cluster, after mid of January and therein the impressions trend have been low for Feb and Mar compared to that of Jan.



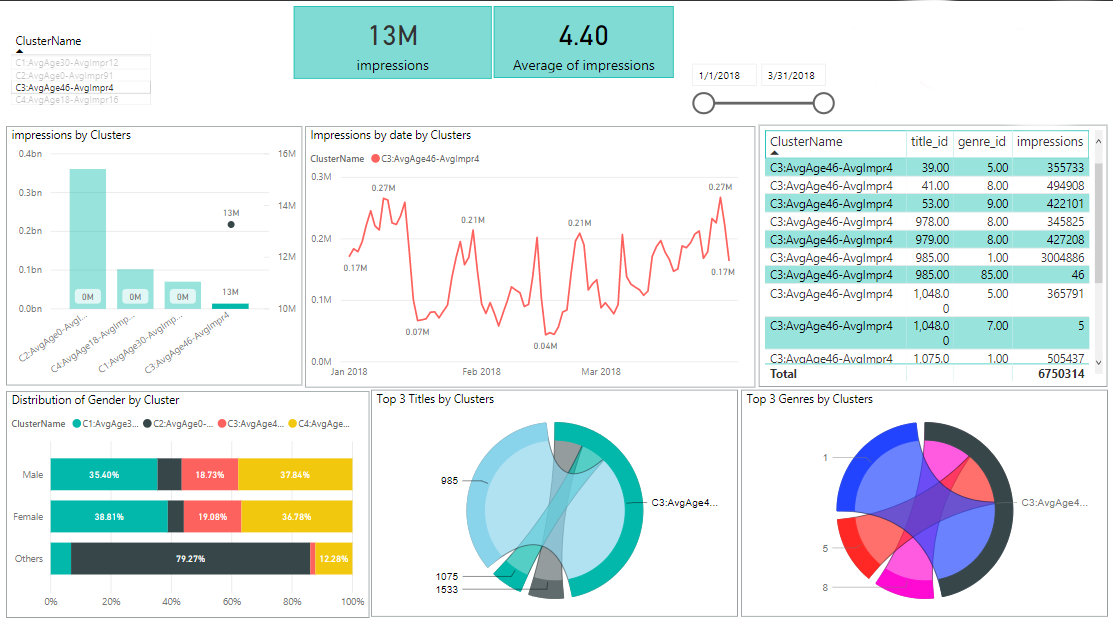
For Cluster C2:

* For this cluster, we see a higher average of impressions(~91 impressions/day)
* In this cluster, top 3 titles are 985, 1533, and 41 which falls in the genres 1, 8 and 5.
* We see steep drops in Impressions, in this cluster, during Jan 13th-Jan 17th, Jan 30th – Feb 2nd, Feb 14th – Feb 18th , Feb 27th – Mar 1st, Mar 12th – Mar 18th, Mar 29th – Mar 31st.
* This clearly shows that there is significant drop in the impressions for this specific cluster, during Mid-months and end of months.



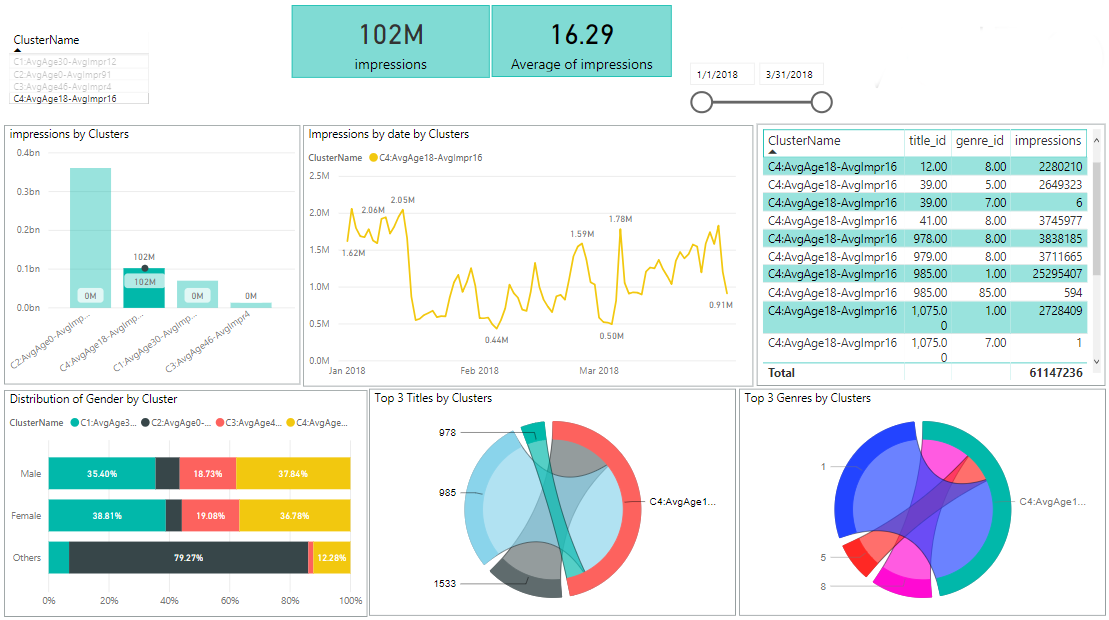
For Cluster C3:

* For this cluster, we see the average of impressions to be ~4 impressions/day, which is the least among all clusters.
* In this cluster, top 3 titles are 985, 1533, and 1075 which falls in the genres 1, 8 and 5.
* In this cluster as well, we see a trend similar to “C2:AvgAge0-AvgImpr91”. We see steep drops in Impressions, in this cluster, during Jan 13th-Jan 17th, Jan 30th – Feb 2nd, Feb 14th – Feb 18th , Feb 27th – Mar 1st, Mar 12th – Mar 18th, Mar 29th – Mar 31st.
* This clearly shows that there is significant drop in the impressions for this cluster, during Mid-months and end of months.



For Cluster C4:

* For this cluster, we see the average of impressions to be ~16 impressions/day.
* In this cluster, top 3 titles are 985, 1533, and 978 which falls in the genres 1, 8 and 5.
* The Impressions trend is similar to that of Cluster 1.



Forecasting was performed on Voiro data using SARIMA(Seasonal ARIMA), to capture any seasonality trends in the data, resulting in a model with ~85% accuracy.

For Age group 20-35, for Male Gender, for the below titles the forecast insights are as follows:

* For Title 41, the forecast model shows a steep increase in impressions from Apr 1st till Apr 5th, followed by a steep drop till Apr 8th, again followed by a steep increase till Apr 12th and a steep drop till by Apr 15th
* For Title 53, the forecast model shows increase in impressions from Apr 2nd till Apr 6th, followed by a drop till Apr 9th, and increase till Apr 12th , ending with a significant drop till Apr 15th.
* For Title 978, the forecast model shows a steep increase in impressions from Apr 2nd till Apr 6th, followed by a steep drop from Apr 6th till Apr 9th, again followed by a steep increase till Apr 12th and a steep drop till Apr 15th.

