

Mentor:

Dr. Narayana D

Final Report

Team members:

* Sulekha Aloorravi
* Deepa Venugopal
* Niharika Thanavarapu
* Lakshmipriya M

**Behavioral pattern recognition of multiplayer online role-playing game players using Big Data Analytics and Machine Learning**

Contents

[I. Domain and Context 3](#_Toc506548784)

[1. Domain 3](#_Toc506548785)

[2. Industry worth 3](#_Toc506548786)

[3. Context of this Project 3](#_Toc506548787)

[4. Objective 3](#_Toc506548788)

[II. Summary 4](#_Toc506548789)

[2.1 Problem Statement 4](#_Toc506548790)

[2.1.1 Perform Exploratory analysis 4](#_Toc506548791)

[2.1.2 Perform Predictive Analytics by applying Machine Learning Models 4](#_Toc506548792)

[2.2 Data 4](#_Toc506548793)

[2.3 Findings and Implications from Dataset 5](#_Toc506548794)

[III. Overview of the final process 6](#_Toc506548795)

[IV. Step-by-step walk through of the solution 7](#_Toc506548796)

[Step 1: Parse Wow Logs 7](#_Toc506548797)

[Step 2: Perform Exploratory Analysis and Clean up data 8](#_Toc506548798)

[Step 3: Perform Exploratory Visualization - Tableau 10](#_Toc506548799)

[Step 4: Perform Feature Engineering (Python and Spark) 15](#_Toc506548800)

[4.1 Create Timeseries Input Features 15](#_Toc506548801)

[4.2 Create Feature Set 1 for Churn Prediction – Method 1 16](#_Toc506548802)

[4.3 Create Feature Set 2 for Churn Prediction – Method 2 18](#_Toc506548803)

[4.4 Create Feature Set 1 for Recommendation Engine – Method 1 21](#_Toc506548804)

[4.5 Create Feature Set 2 for Recommendation – Method 2 23](#_Toc506548805)

[Step 5: Perform Model Selection and Tuning (Python and Spark) 24](#_Toc506548806)

[5.1 KMeans Clustering for Data Exploration 24](#_Toc506548807)

[5.2 Time Series Forecasting of hourly data using Holt Winters model 26](#_Toc506548808)

[5.3 Time Series Forecasting of hourly data using ARIMA model 29](#_Toc506548809)

[5.4 Time Series Forecasting of hourly data using LSTM Tensowflow 31](#_Toc506548810)

[5.5 Churn Prediction – Method 1 on Feature Set 1 32](#_Toc506548811)

[5.6 Churn Prediction – Method 2 on Feature Set 2 36](#_Toc506548812)

[5.7 Recommendation Engine – KNN and User Based Collaborative Engine – Feature Set 1 41](#_Toc506548813)

[5.8 Recommendation Engine – KNN and User Based Collaborative Engine – Feature Set 2 43](#_Toc506548814)

[V. Model Evaluation 44](#_Toc506548815)

[i. Forecast the number of players expected in future time point 44](#_Toc506548816)

[Holt-Winters seasonal method 45](#_Toc506548817)

[ii. Predict player churning 46](#_Toc506548818)

[Decision Tree Classifier 47](#_Toc506548819)

[iii. Recommend Guild, Race/Class to players 47](#_Toc506548820)

[KNN and User based collaborative filtering 47](#_Toc506548821)

[VI. Comparison to benchmark 48](#_Toc506548822)

[Comparison of Final solution to initial benchmark 49](#_Toc506548823)

[VII. Visualizations 49](#_Toc506548824)

[Cluster Data Analysis 49](#_Toc506548825)

[Churn Data Analysis 52](#_Toc506548826)

[WOW Zone Heat map 54](#_Toc506548827)

[VIII. Implications 55](#_Toc506548828)

[IX. Limitations 55](#_Toc506548829)

[X. Closing Reflections 55](#_Toc506548830)

[XI. Code Information 56](#_Toc506548831)

[Appendix 58](#_Toc506548832)

[Other models attempted before interim report 58](#_Toc506548833)

[References 70](#_Toc506548834)

# Domain and Context

## Domain

A massively multiplayer online game (more commonly, MMO) is an online game which is capable of supporting large numbers of players, typically from hundreds to thousands, simultaneously from around the world.

These games can be found for most network-capable platforms, including the personal computer, video game console, or smartphones and other mobile devices. MMOs can enable players to cooperate and compete with each other on a large scale, and sometimes to interact meaningfully with people around the world.

## Industry worth

The UK MMO-market is worth £195 million in 2009 compared to the £165 million and £145 million spent by German and French online gamers. The US gamers spend more, however, spending about $3.8 billion overall on MMO games. $1.8 billion of that money is spent on monthly subscription fees. The money spent averages out to $15.10 between both subscription and free-to-play MMO gamers. The study published by “Today’s Gamers MMO Focus Report” also found that 46% of 46 million players in the US pay real money to play MMO games.

## Context of this Project

It is challenging to develop the database engines that are needed to run a successful MMOG with millions of players. Understanding the behavior of players using their activity data is more important for these game developers to come up with better strategies in game development.

The variety, volume, velocity, value and veracity (Big Data 5Vs) of data that is involved in these Gaming environments exceed the limits of analysis and manipulation of conventional tools, therefore, Big Data platforms are required to handle and interpret this data.

Great volumes of data are generated all the time in these environments. Each interaction made by a player creates data that are transferred and stored, and if properly analyzed, can contain valuable information. This information can be vital for the continuity and improvement of a game. Patterns can be detected from these data and even predictive analysis can be made to foresee the actions and intentions of the players inside the game.

## Objective

Objective of this Project is to perform analytics on one such Big Data Gaming Environment and the results would help game developers in:

* Optimizing user experience
* Improving revenue
* Raise the level of control over the environment

# Summary

## 2.1 Problem Statement

### 2.1.1 Perform Exploratory analysis

* 1. To cluster players into different groups based on features in dataset
  2. To analyze and visualize timeline patterns of players by different groups and parameters
  3. To create heat map based on the gaming zones
  4. To visualize patterns based on Guilds they belong to

### 2.1.2 Perform Predictive Analytics by applying Machine Learning Models

* 1. **Forecast the number of players expected in future time point** to plan resource capacity
  2. **Predict player churning** to come up with steps to avoid future churn
  3. **Recommend guilds** [groups to join] **to players** for effective gaming and to minimize churn

## 2.2 Data

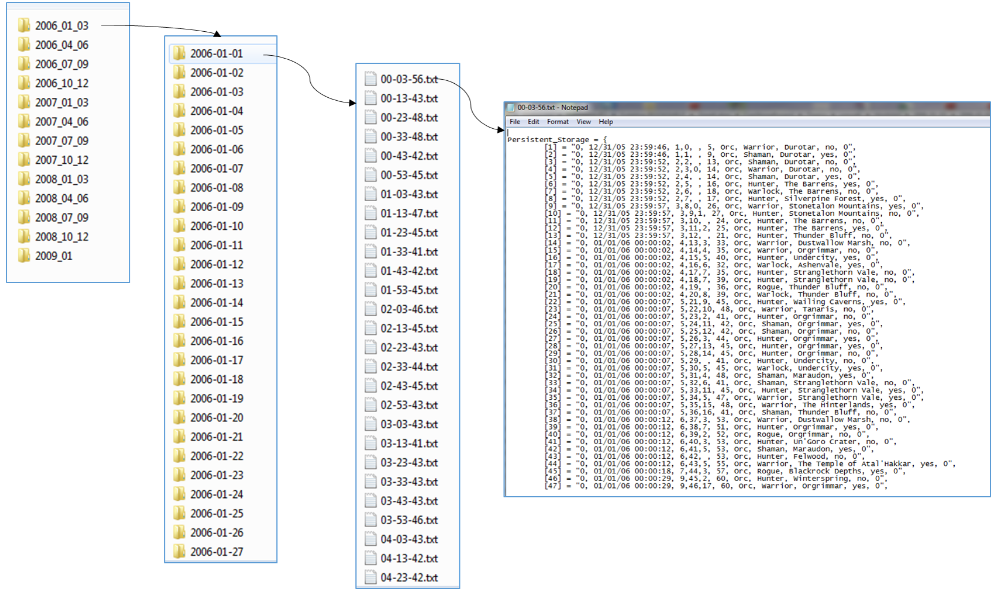
We have chosen an online game named “World of Warcraft” which we found to be most suitable for this Project. A large and scalable dataset with 3 years of player logs are released by Blizzard Entertainment for research purposes. We are using this dataset of our Project.

|  |  |
| --- | --- |
| **Data set Summary** | |
| **Attribute** | **Value** |
| Data duration (in days) | 1107 |
| Sampling Rate per day | 124 |
| No. of Samples | 138084 |
| No. of Records (rows) | 36,513,647 |
| No. of Values (Data points) | 438,163,764 |
| Size of data (in GB) | 3.4 |
| Dataset Type | Logs |
| Format | Text Files |
| No. of Folders | 1095 |

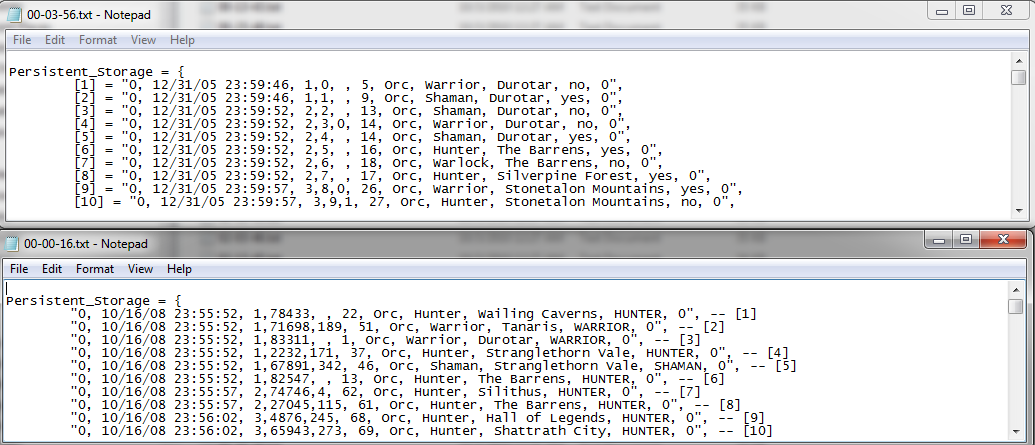
|  |  |  |
| --- | --- | --- |
| **Field Description** | | |
| **Field** | **Description** | **Data Type** |
| Query Time | Date and time when logs were generated | integer |
| Query Seq. # | Sequence of queries | integer |
| Avatar ID | Unique id for each user | integer |
| Guild | Group id of the player | integer |
| Level | Game level of the player | integer |
| Race | Blood Elf, Orc, Tauren, Troll, Undead | String |
| Class | Death Knight, Druid, Hunter, Mage, Paladin, Priest, Rogue, Shaman, Warlock, Warrior | String |
| Zone | One of the 229 Zones in World of WarCraft game | String |

## 2.3 Findings and Implications from Dataset

1. Dataset is in the following folder and file structure as shown in the below image



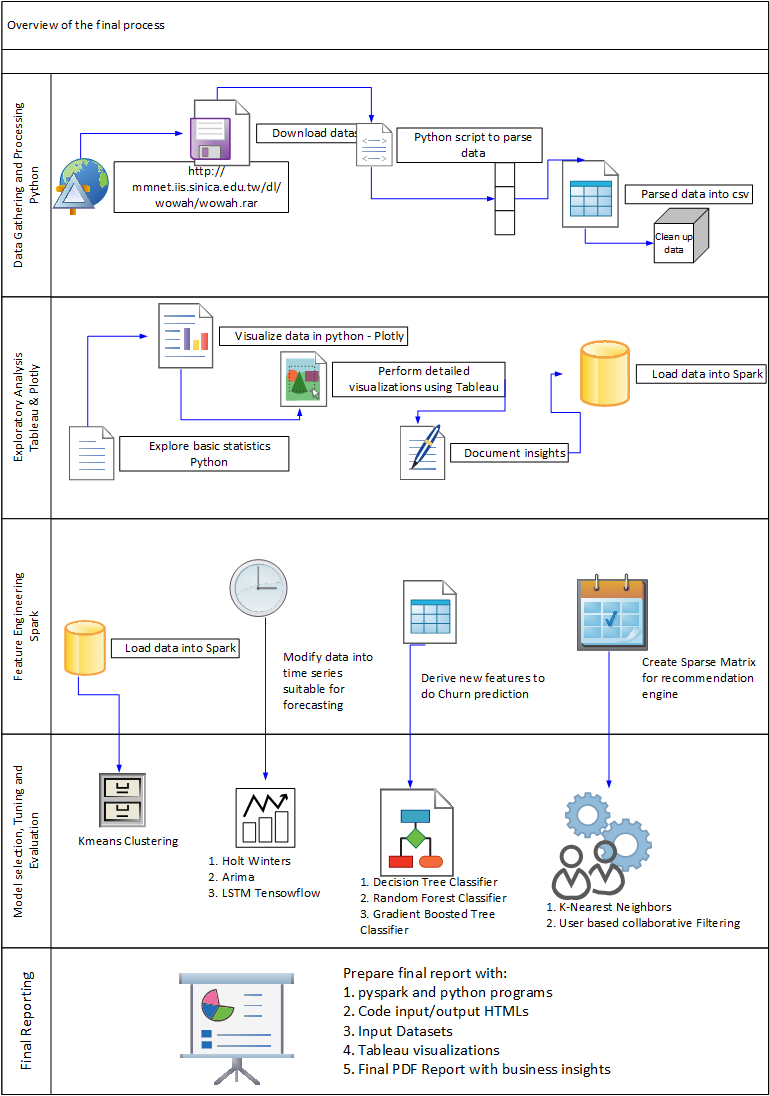
1. There are two different types of representations of logs within the folders



1. Both the types of logs need to be parsed and collected into a single csv file after removing irrelevant information.
2. Fortunately there are no data losses within the 3 year period of 2006 to 2009

# Overview of the final process

Described below is the detailed process flow with algorithms and techniques used :



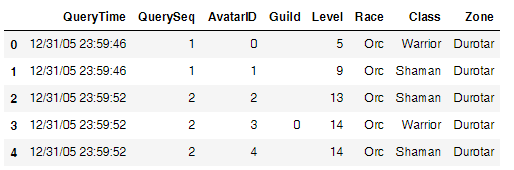
# Step-by-step walk through of the solution

## Step 1: Parse Wow Logs

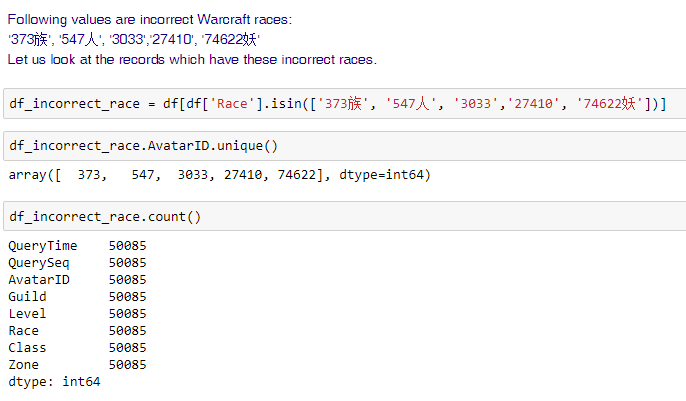


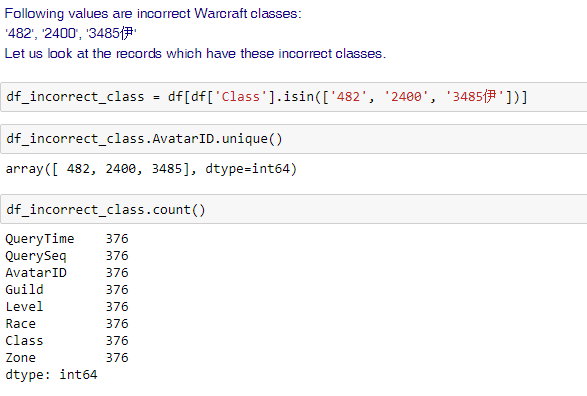


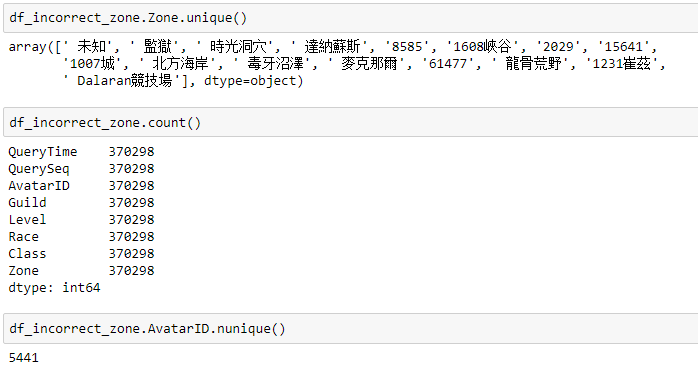


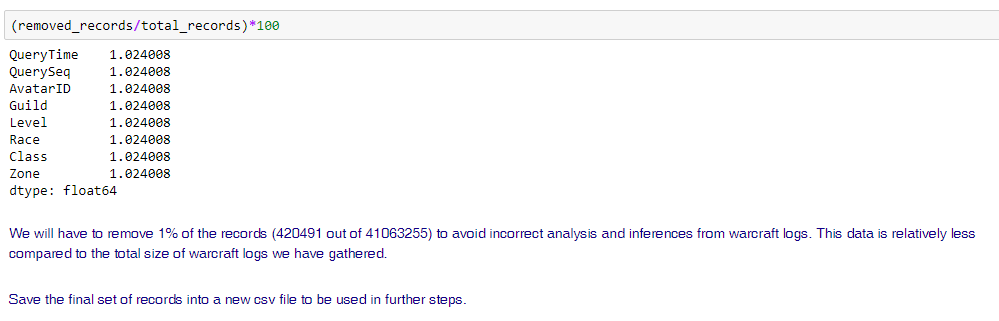


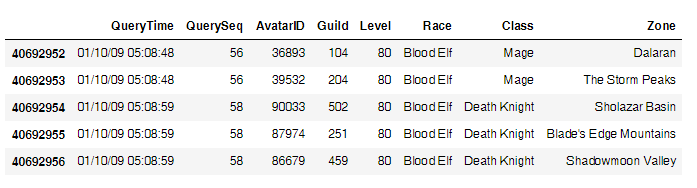
## Step 2: Perform Exploratory Analysis and Clean up data





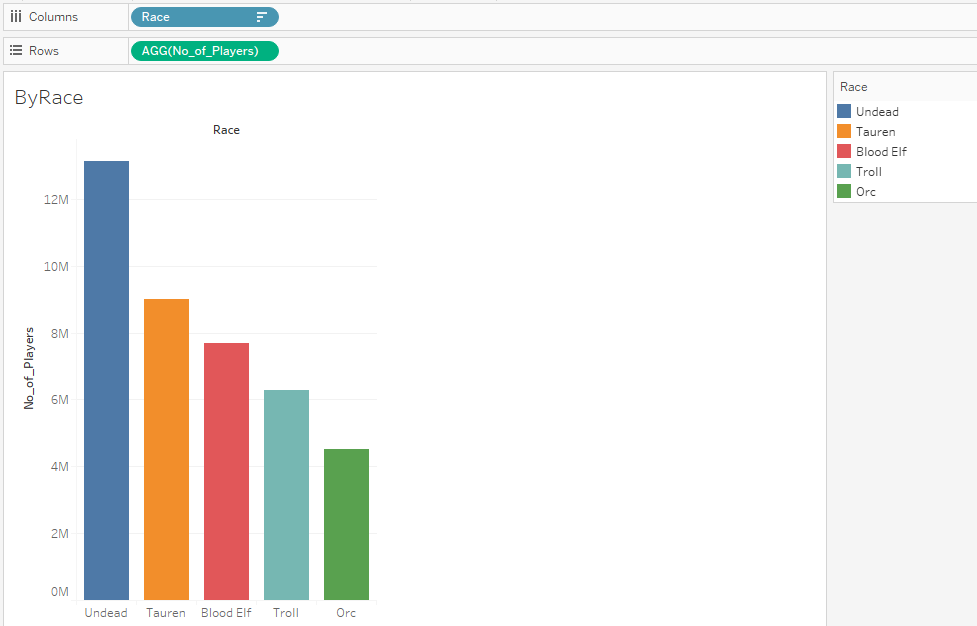


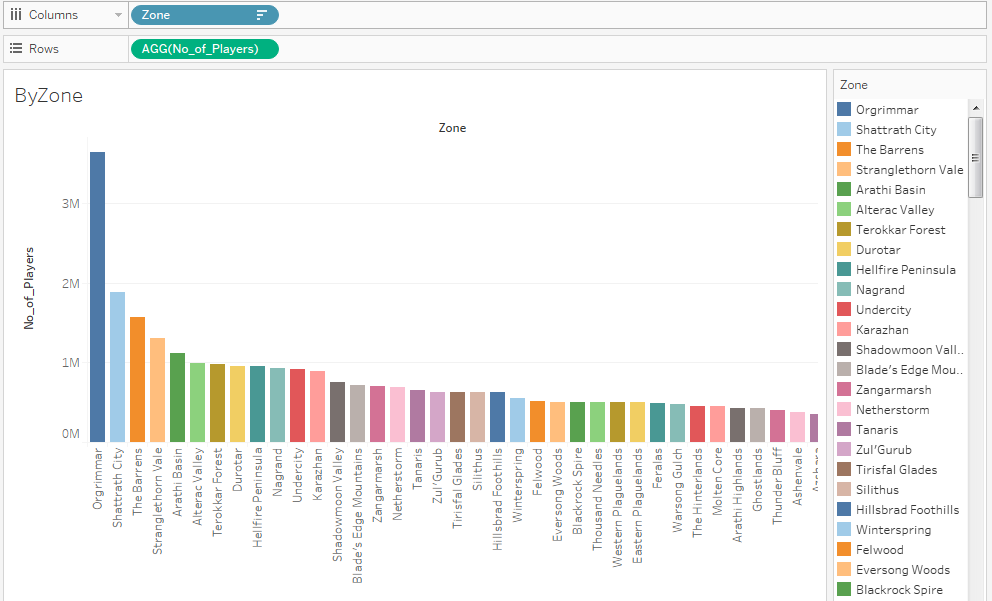


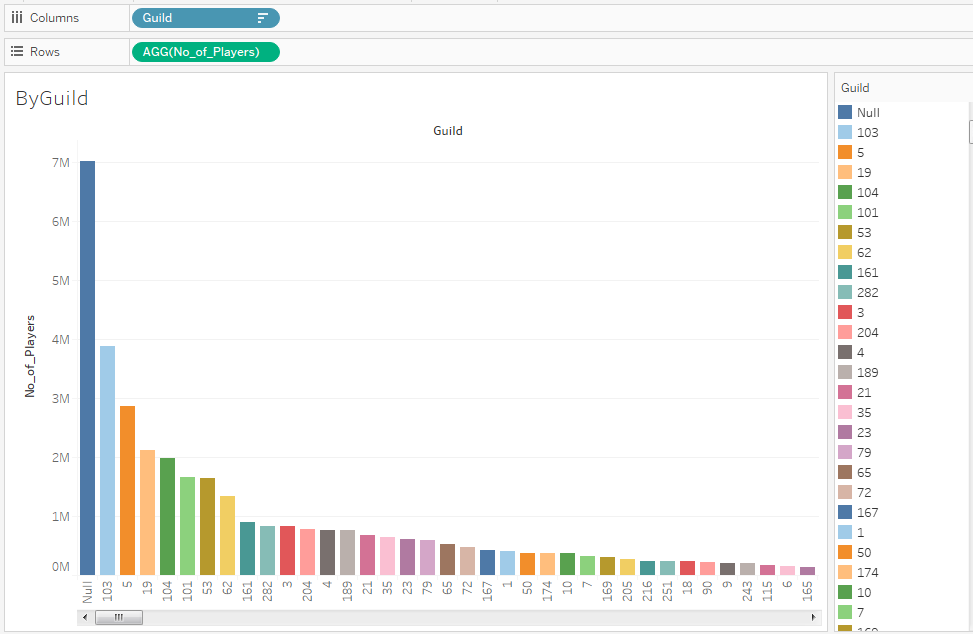


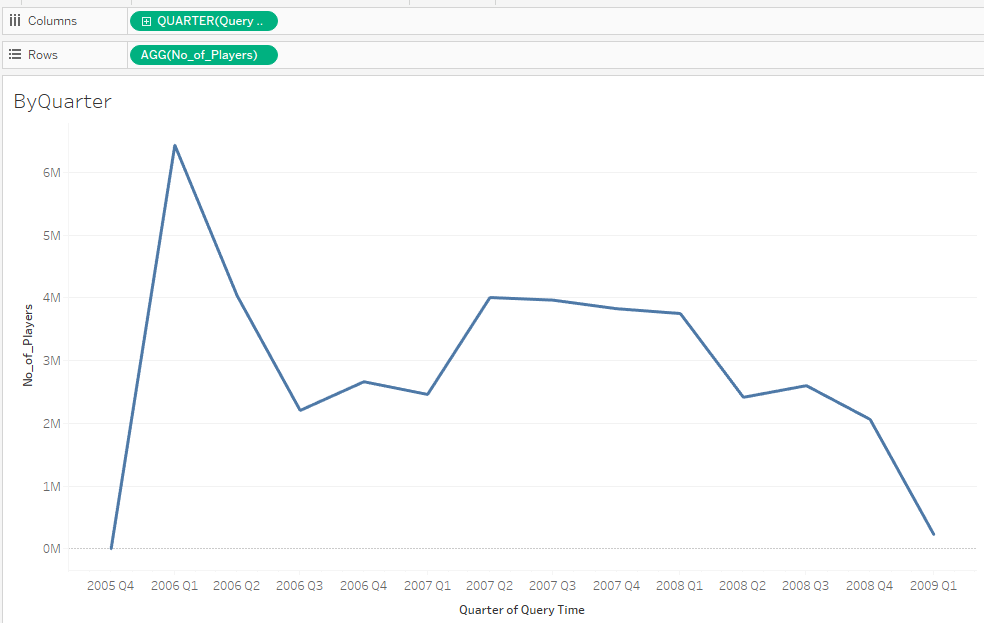
## Step 3: Perform Exploratory Visualization - Tableau



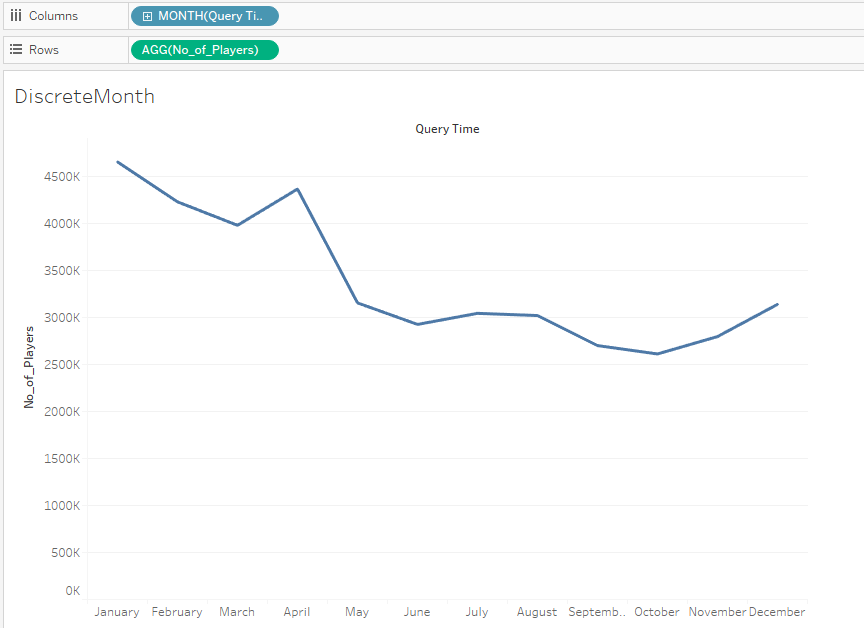


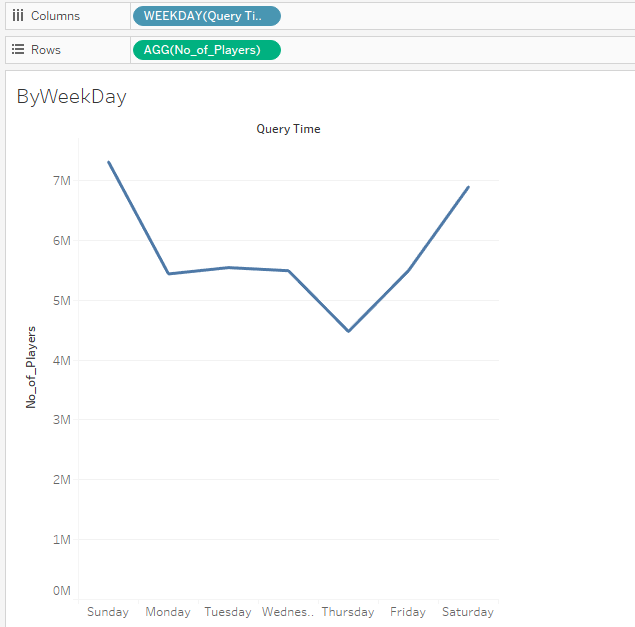


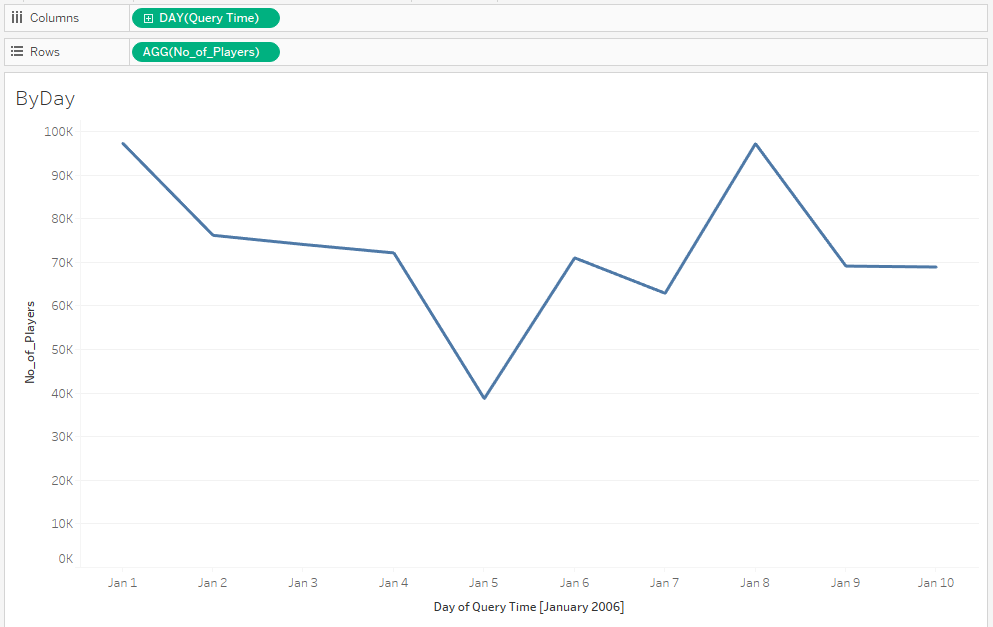


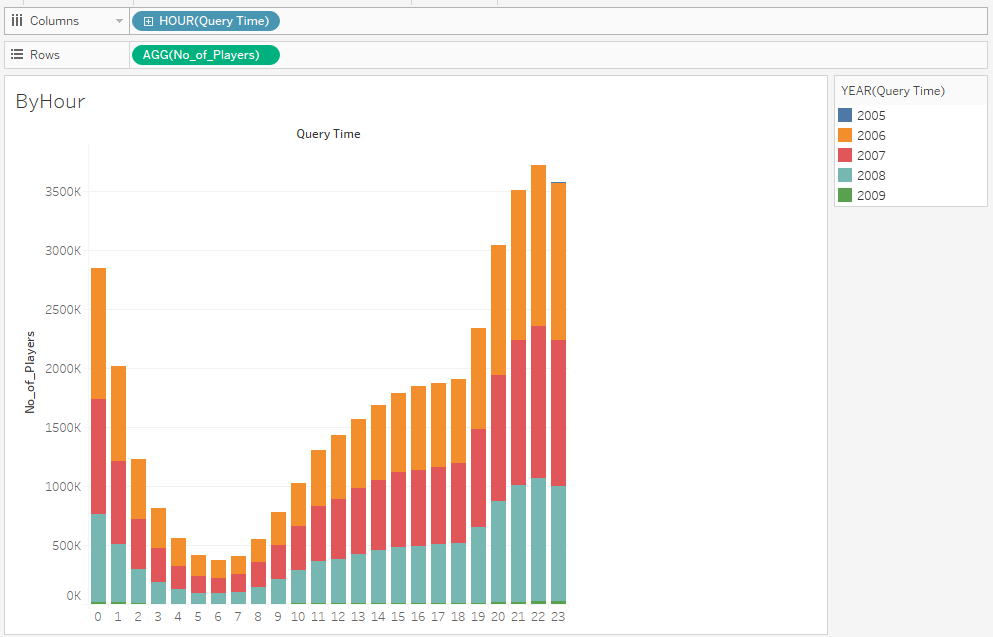








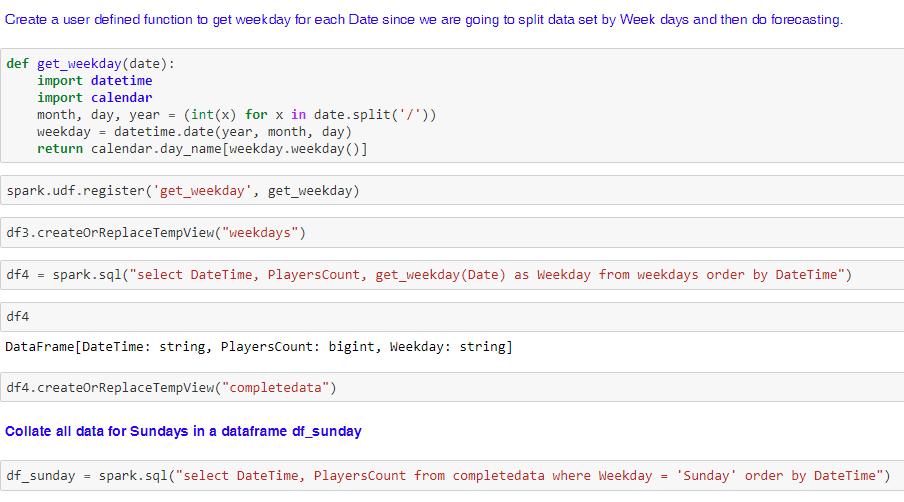


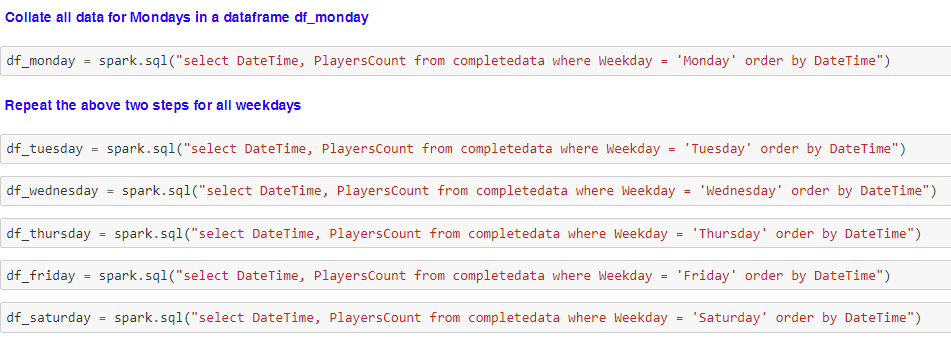


## Step 4: Perform Feature Engineering (Python and Spark)

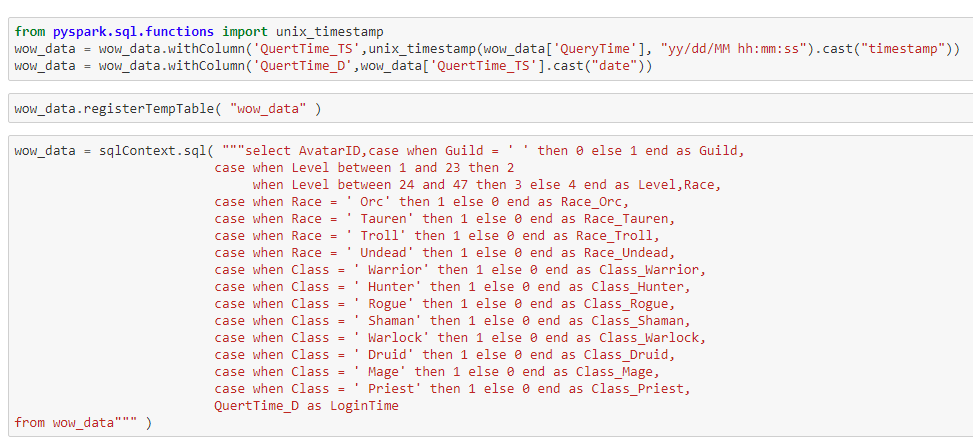
### Create Timeseries Input Features

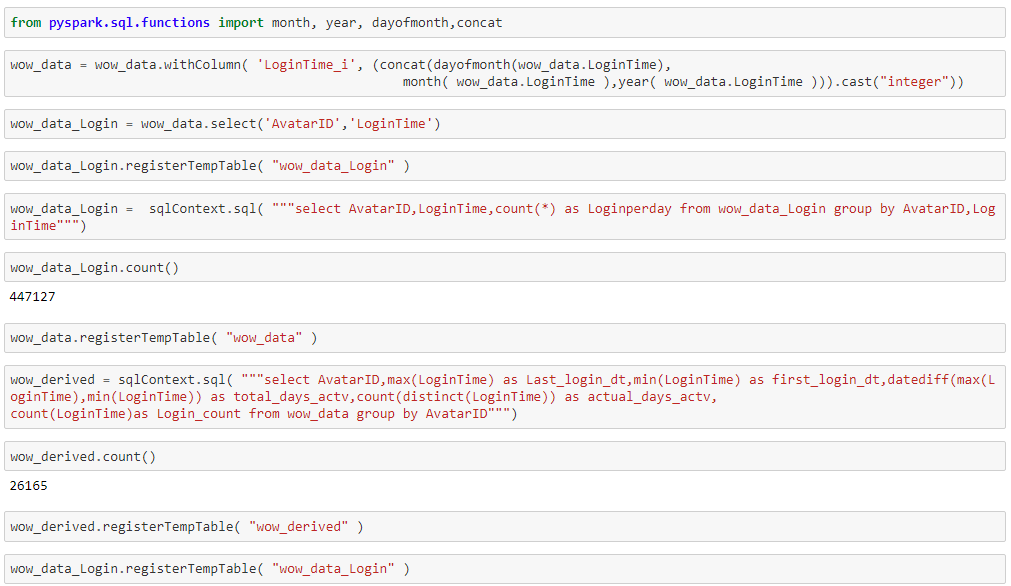


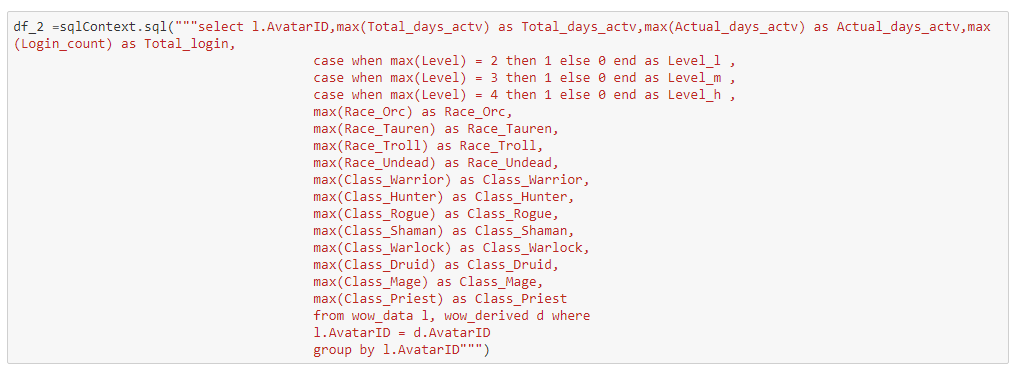




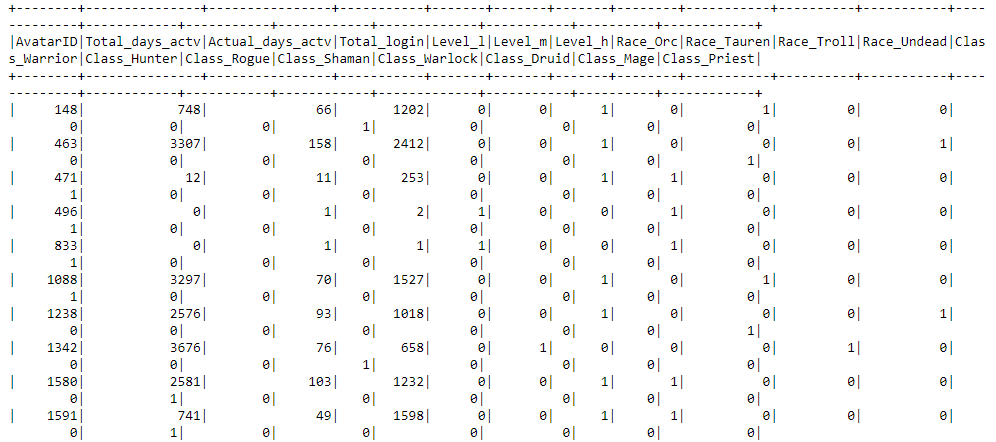
### Create Feature Set 1 for Churn Prediction – Method 1



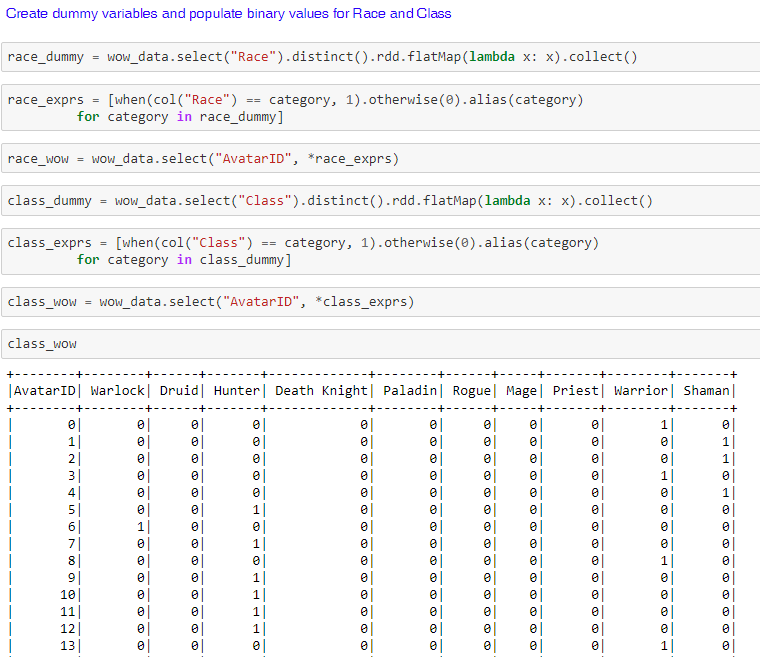




**New Features**

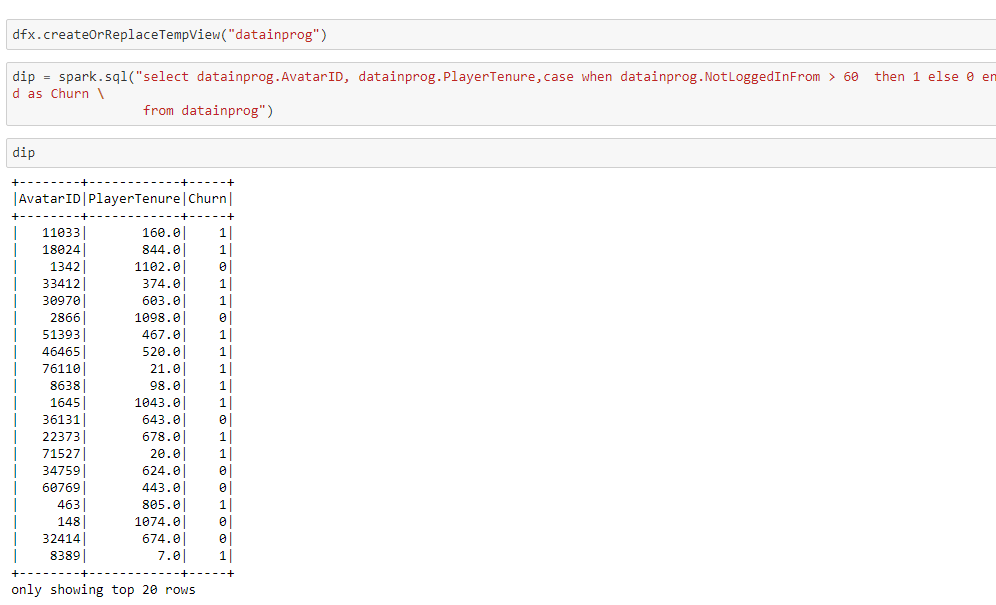
****

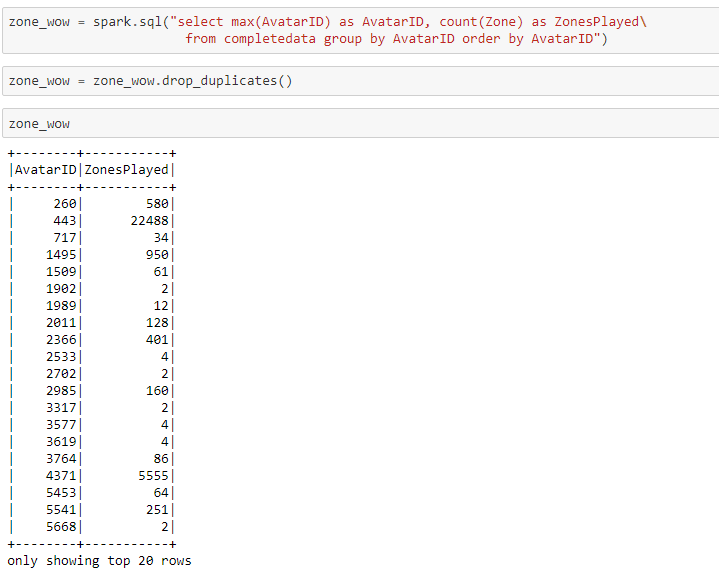
### Create Feature Set 2 for Churn Prediction – Method 2



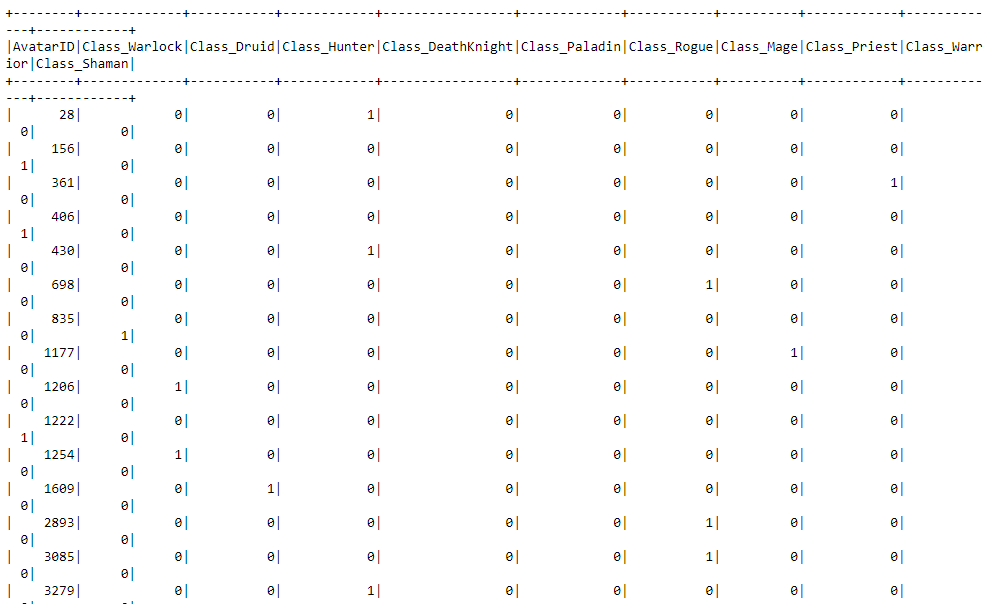


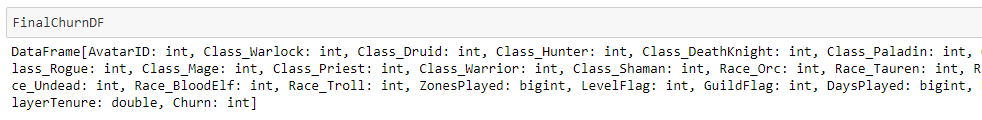






**New Features**

****

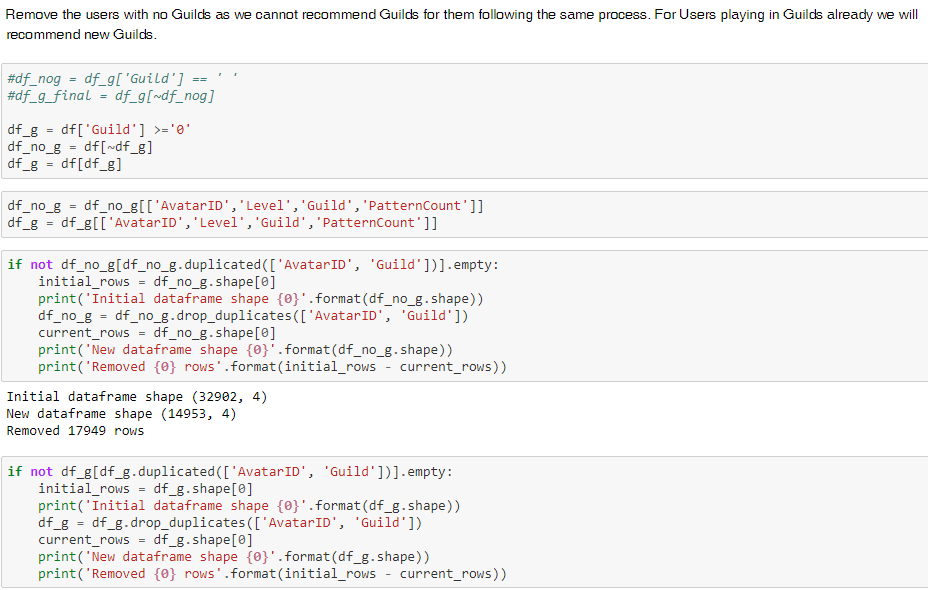
****

### Create Feature Set 1 for Recommendation Engine – Method 1



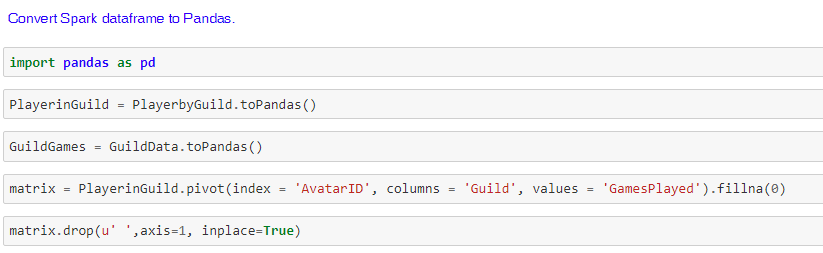






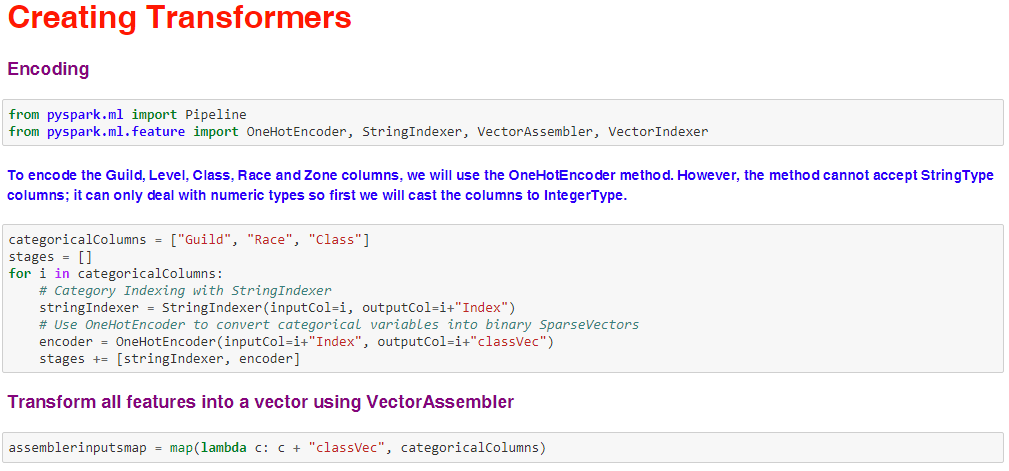
### Create Feature Set 2 for Recommendation – Method 2

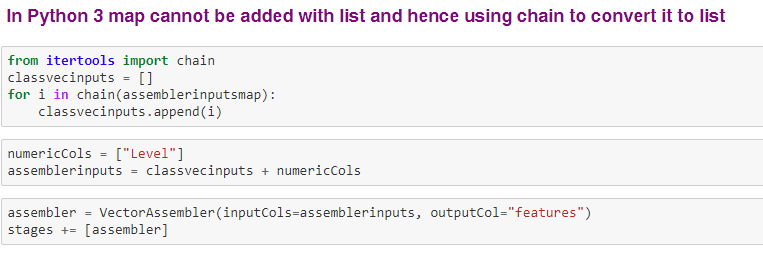


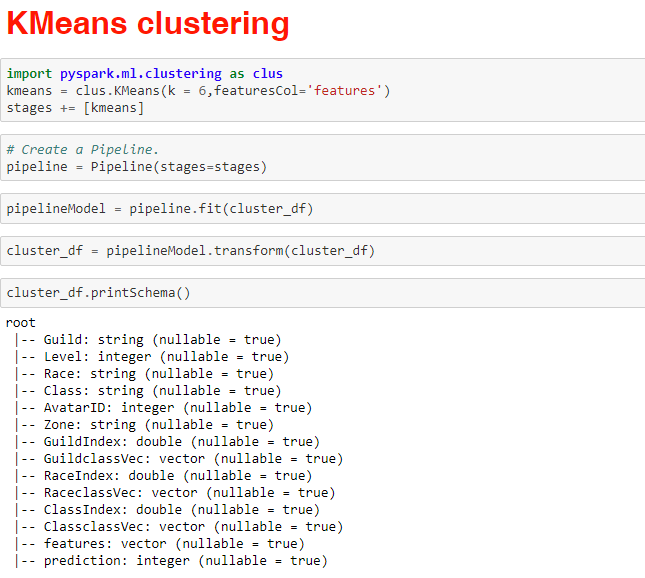


## Step 5: Perform Model Selection and Tuning (Python and Spark)

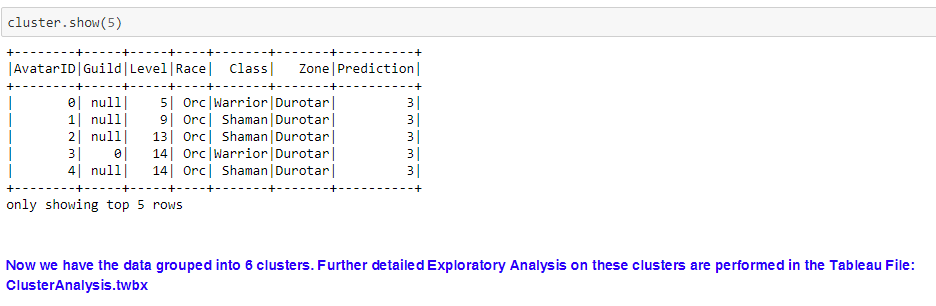
### 5.1 KMeans Clustering for Data Exploration



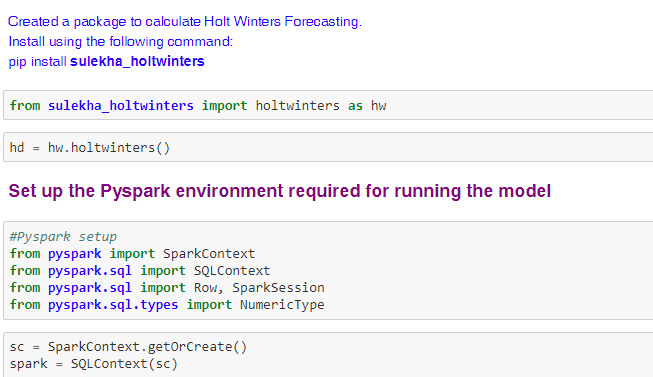


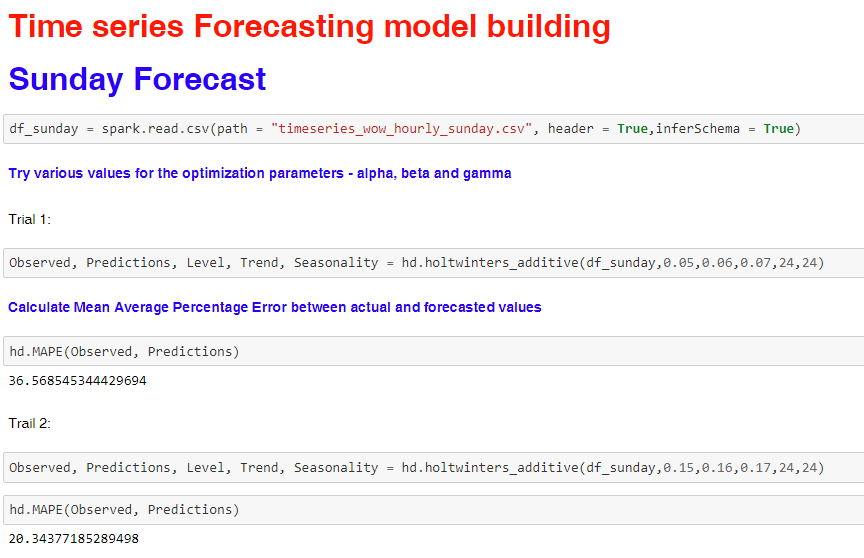


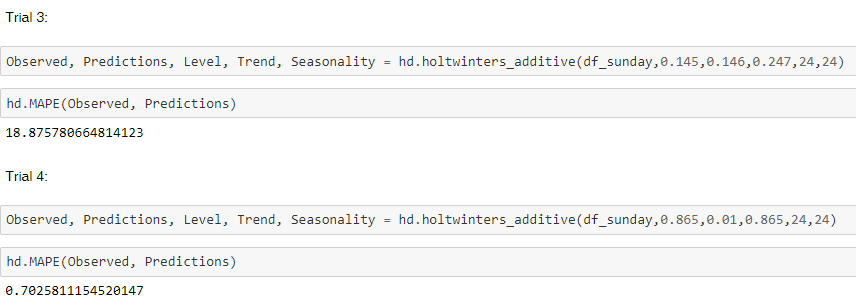


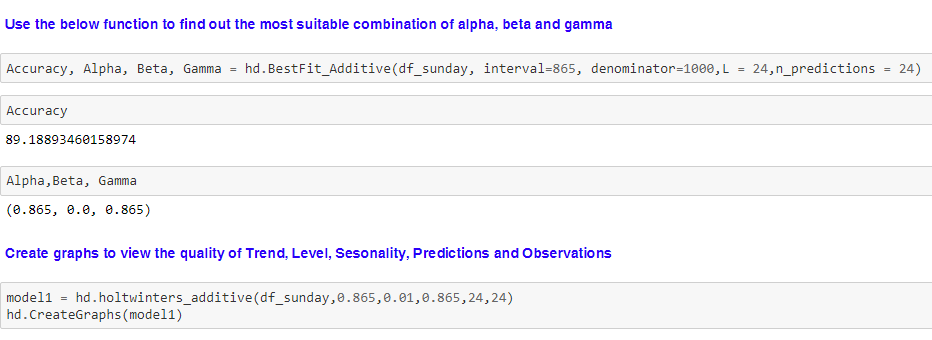


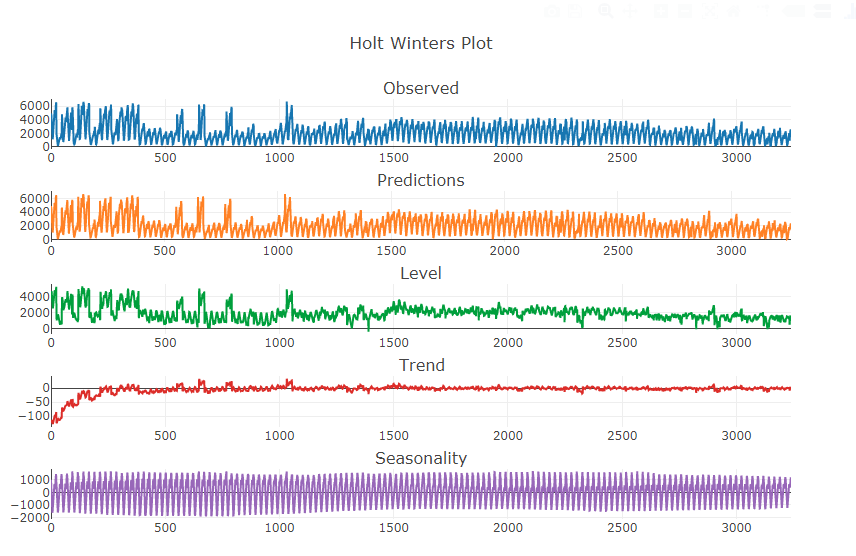
### 5.2 Time Series Forecasting of hourly data using Holt Winters model

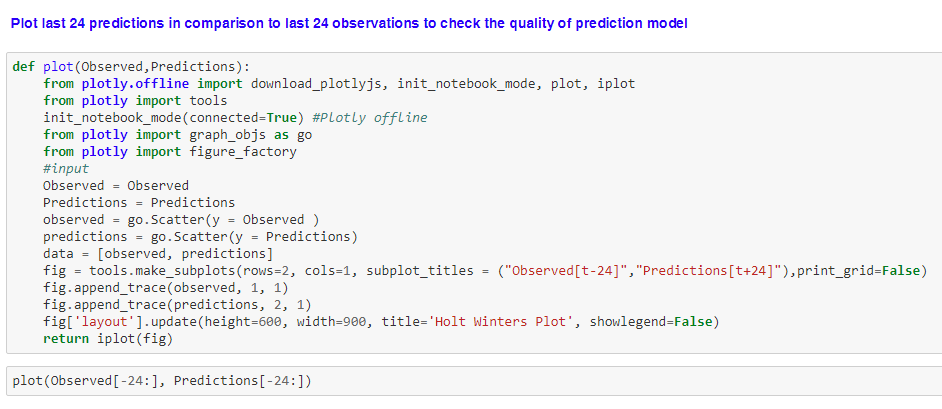


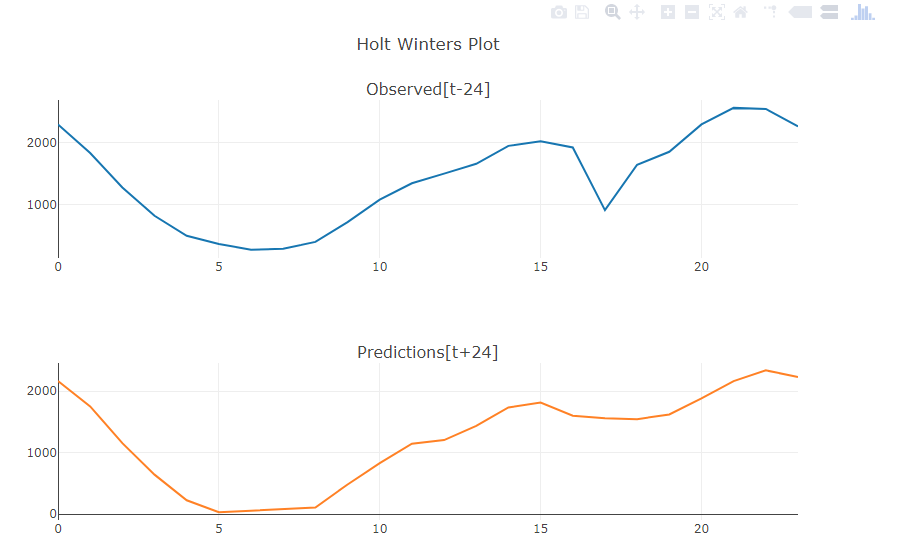




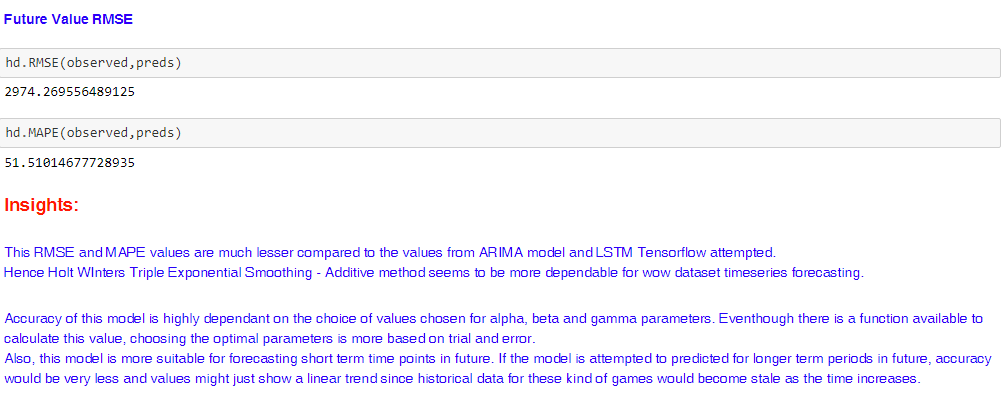




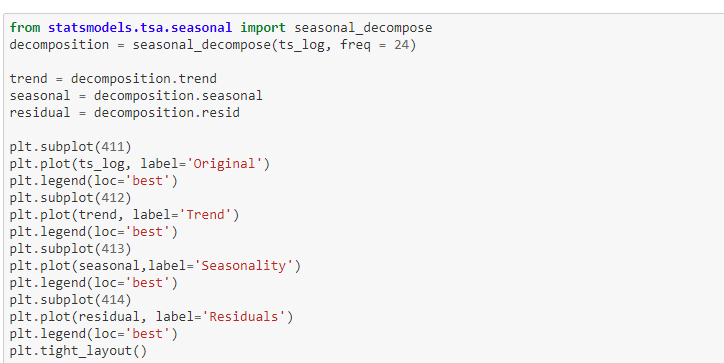


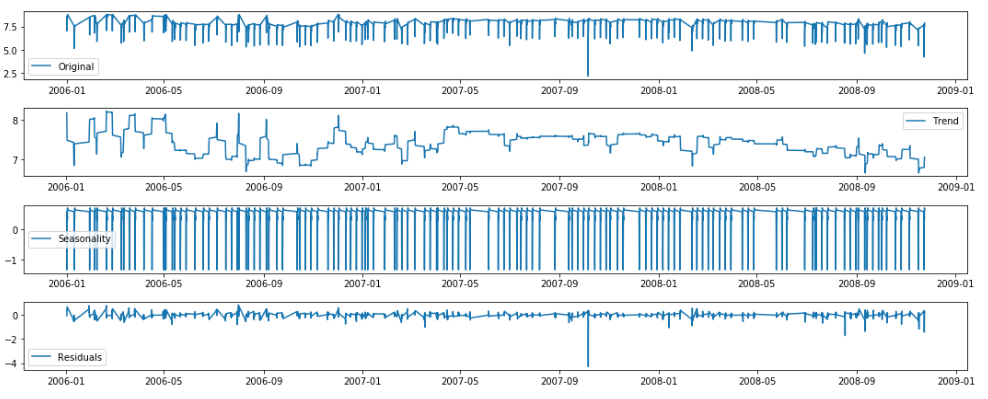


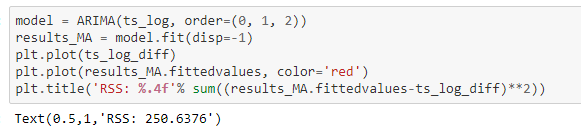


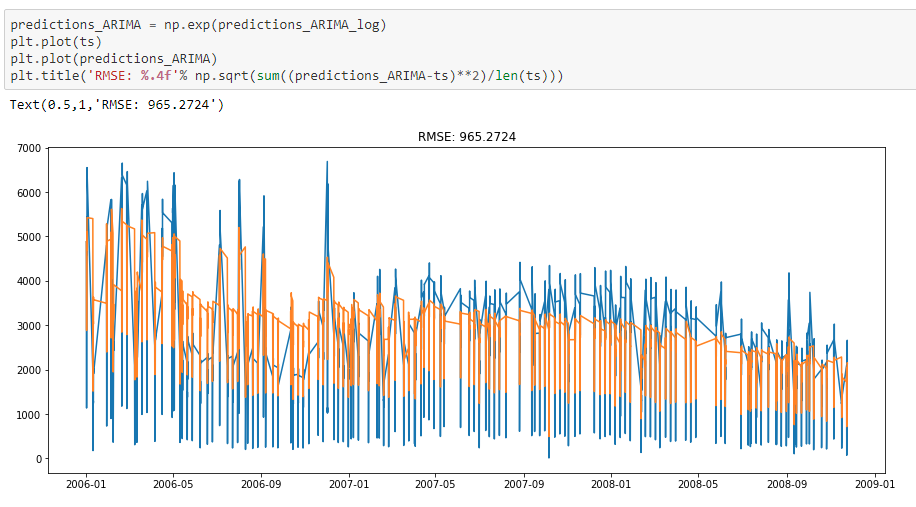


### 5.3 Time Series Forecasting of hourly data using ARIMA model



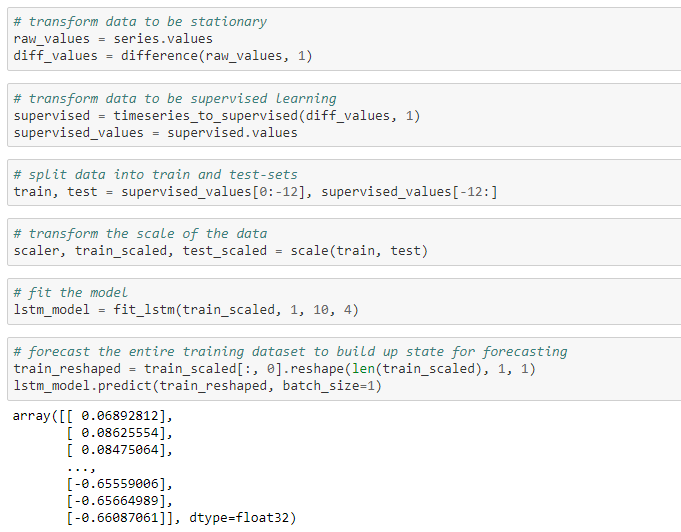


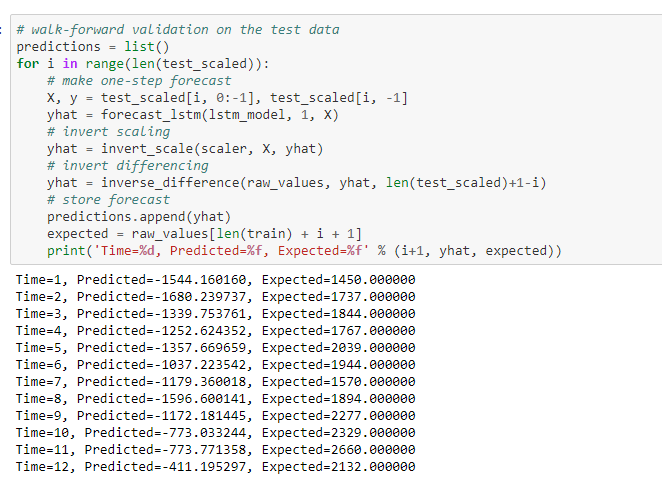


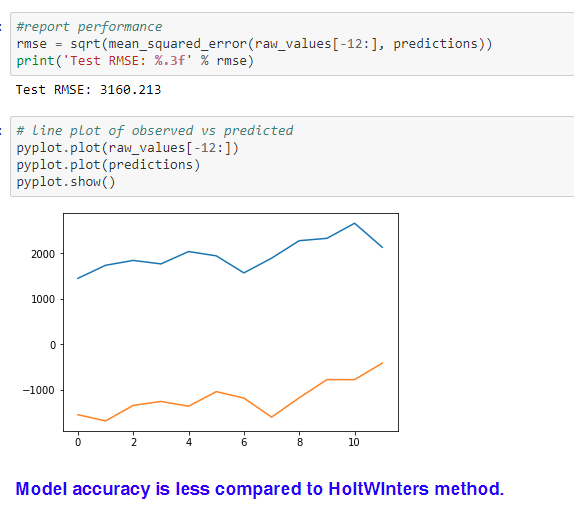




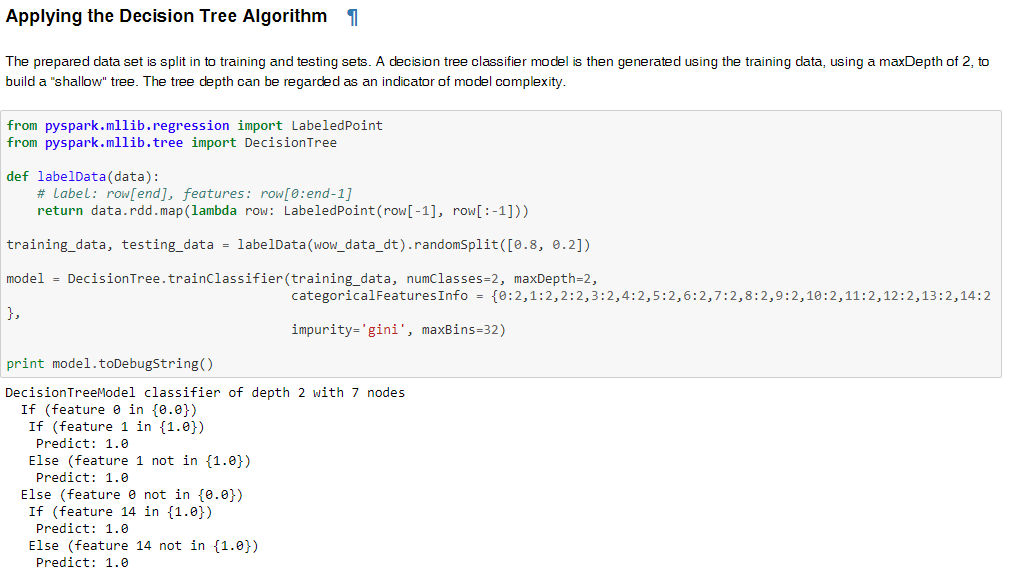
### 5.4 Time Series Forecasting of hourly data using LSTM Tensowflow

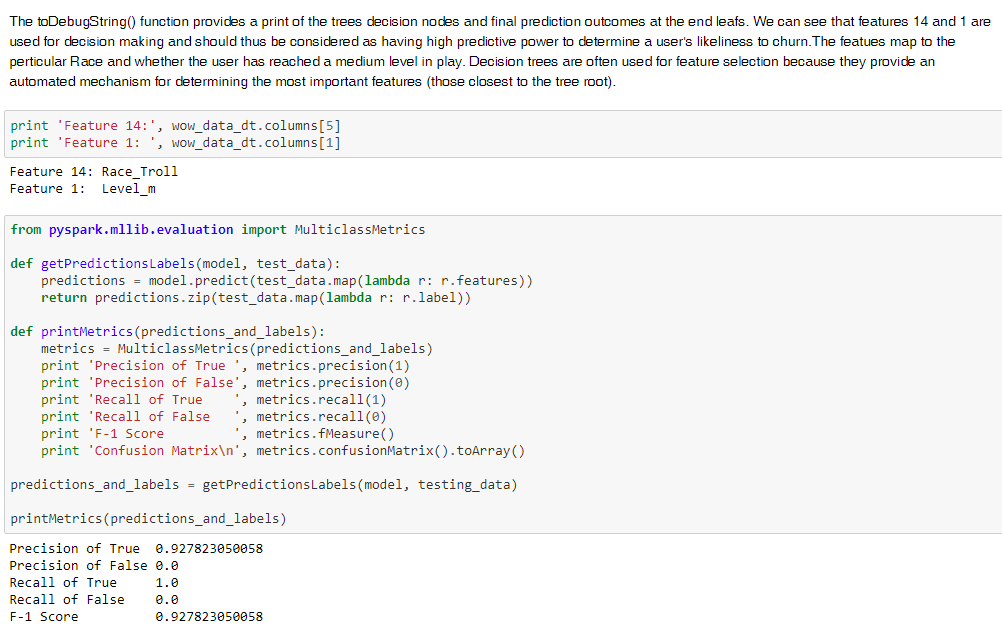


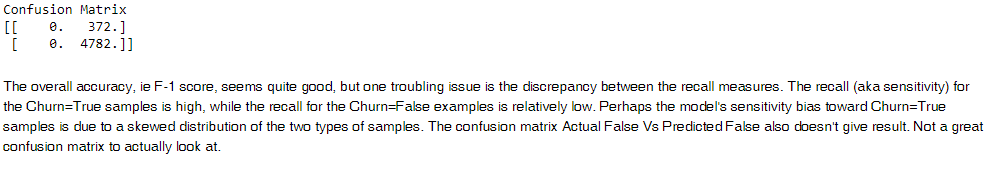


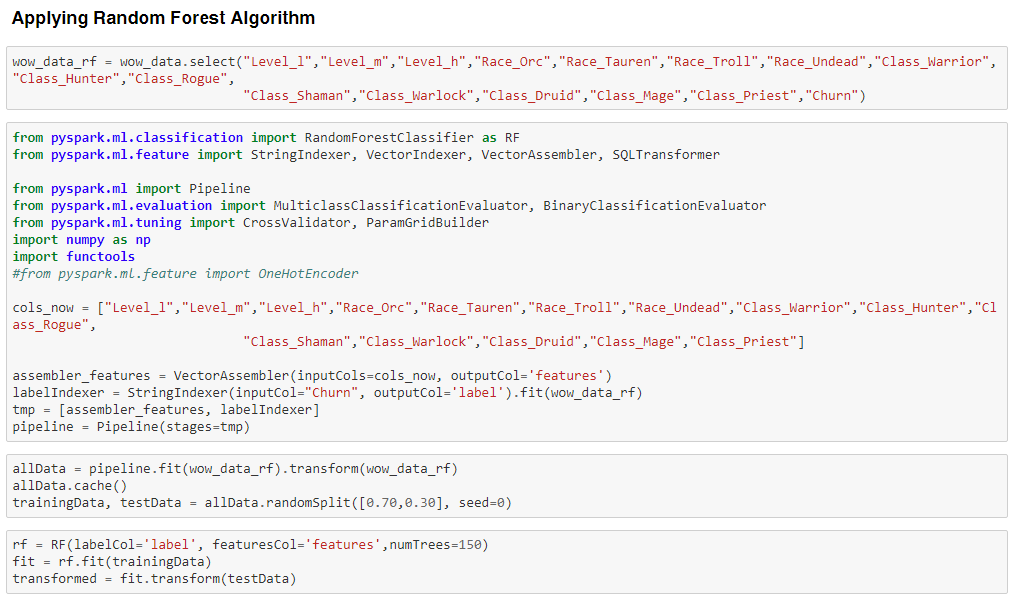


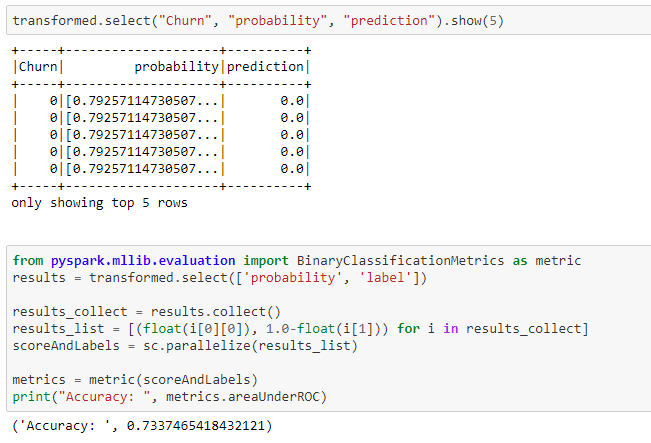
### 5.5 Churn Prediction – Method 1 on Feature Set 1



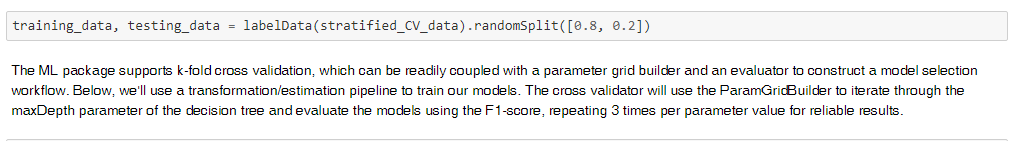


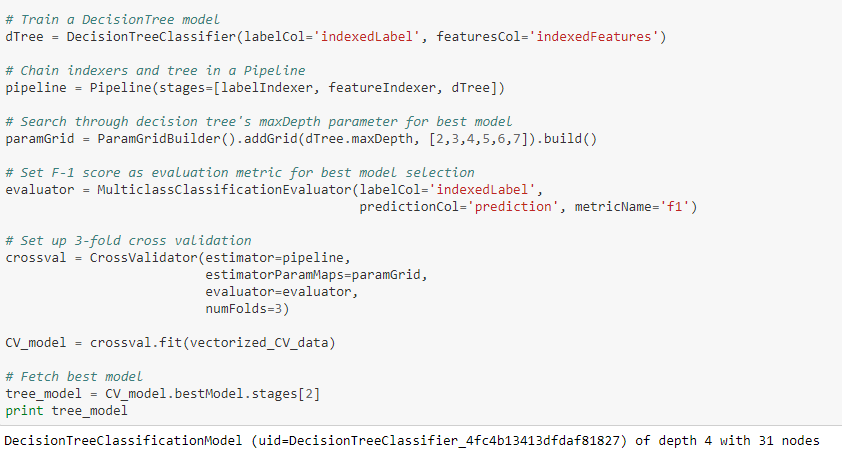


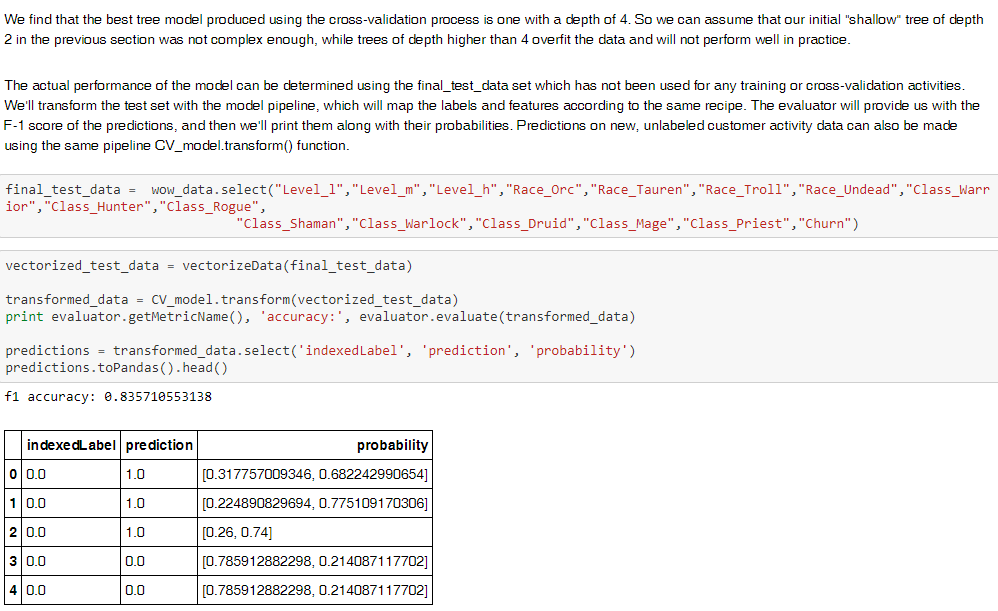








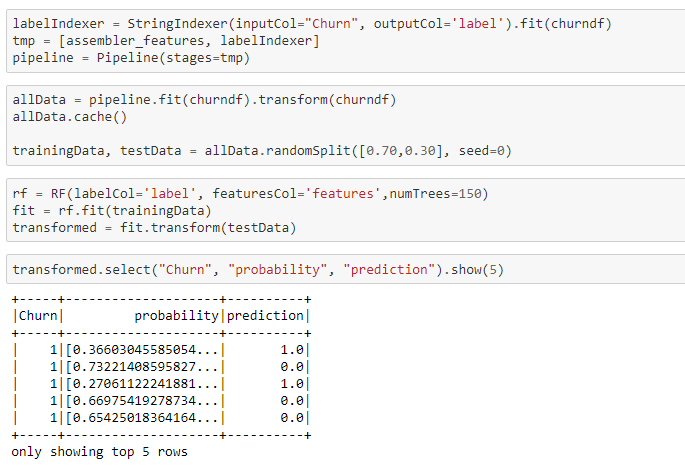


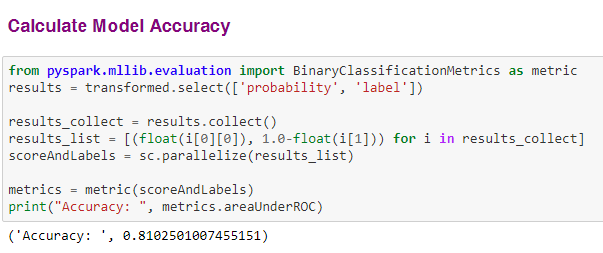


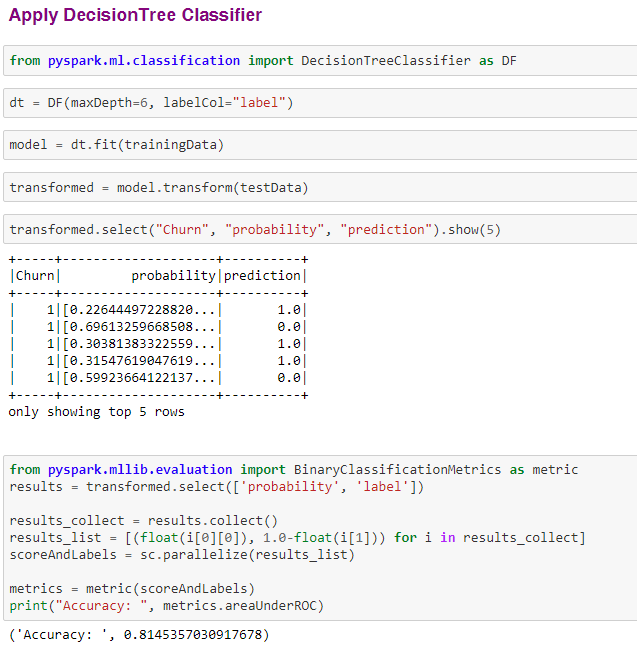


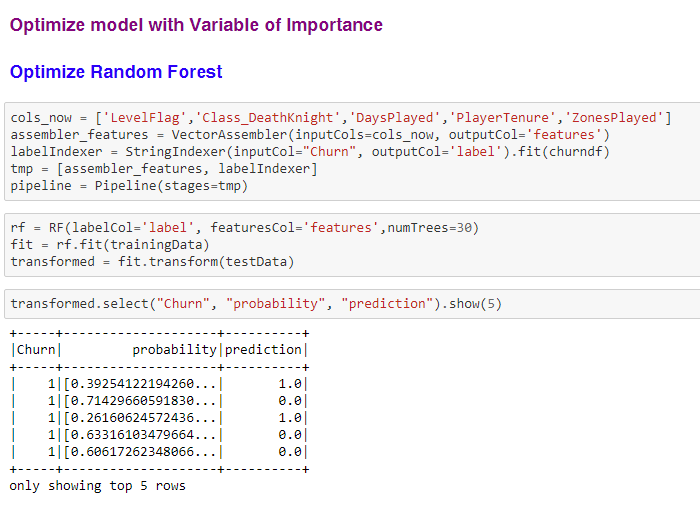
### 5.6 Churn Prediction – Method 2 on Feature Set 2

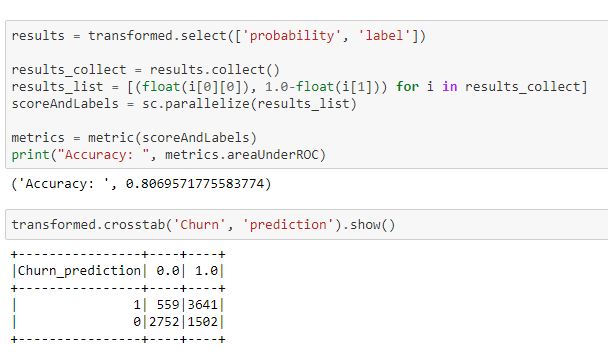


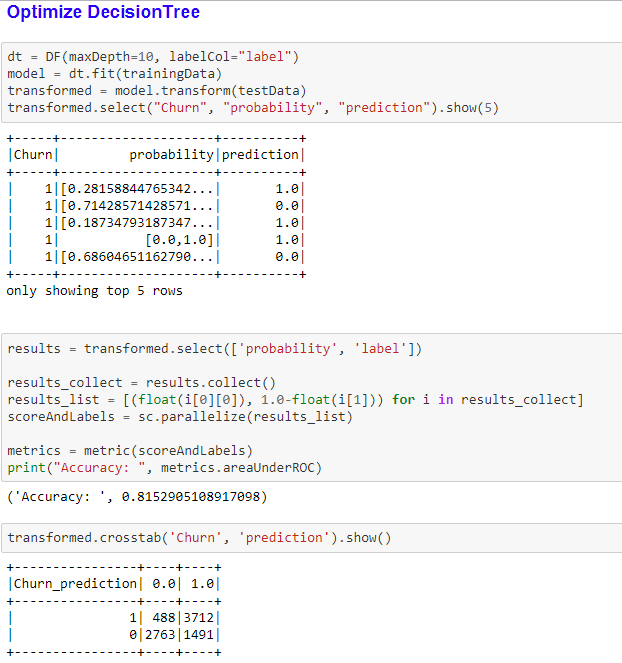


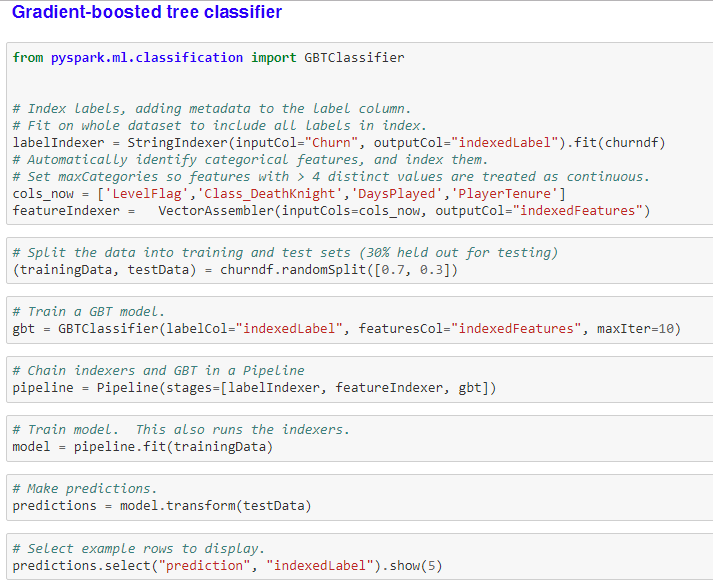


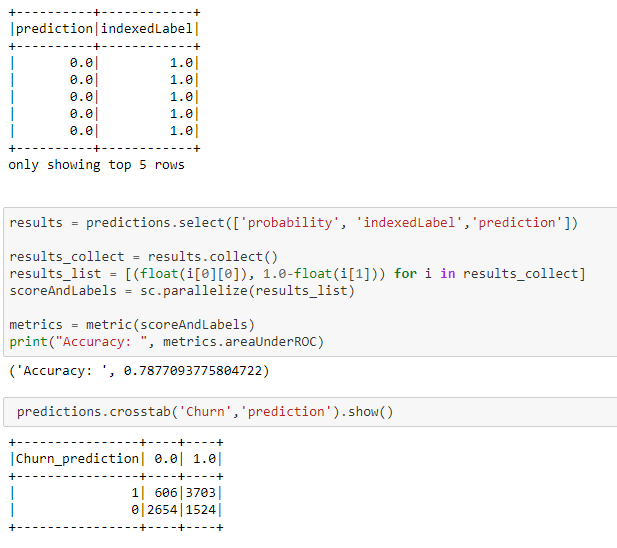


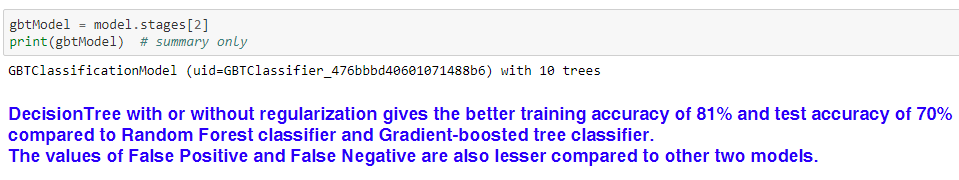




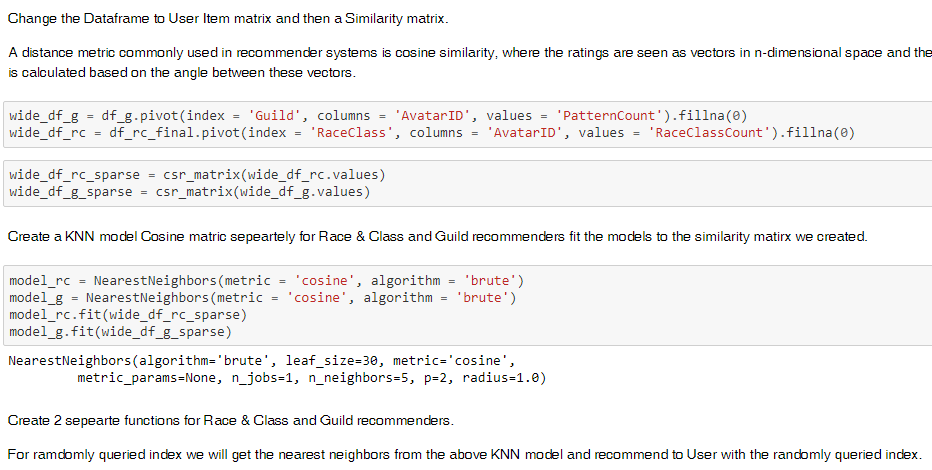




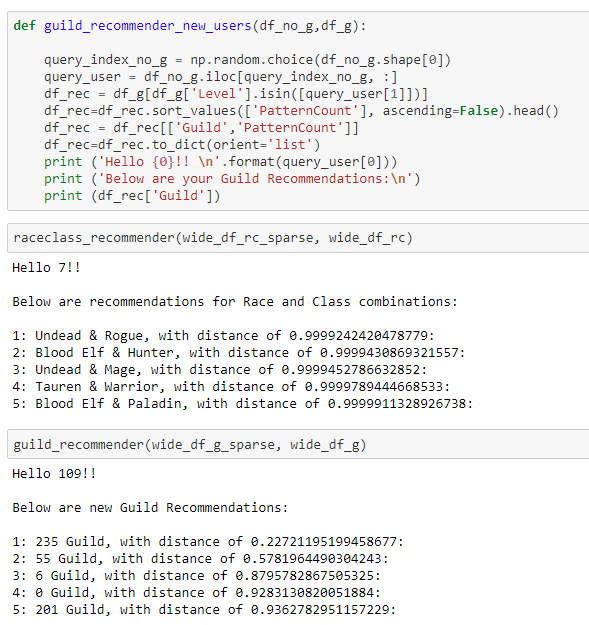


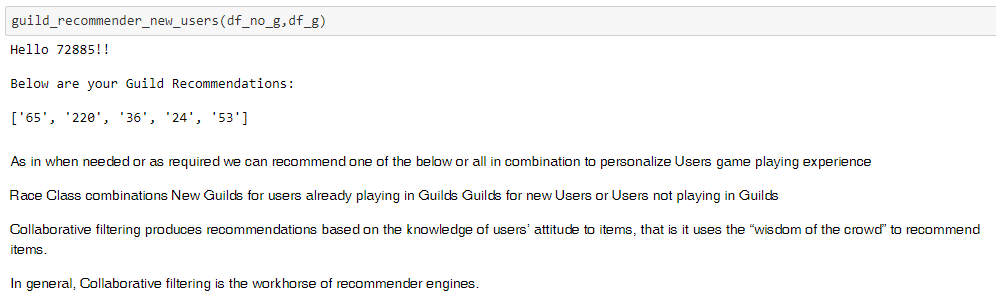


### 5.7 Recommendation Engine – KNN and User Based Collaborative Engine – Feature Set 1

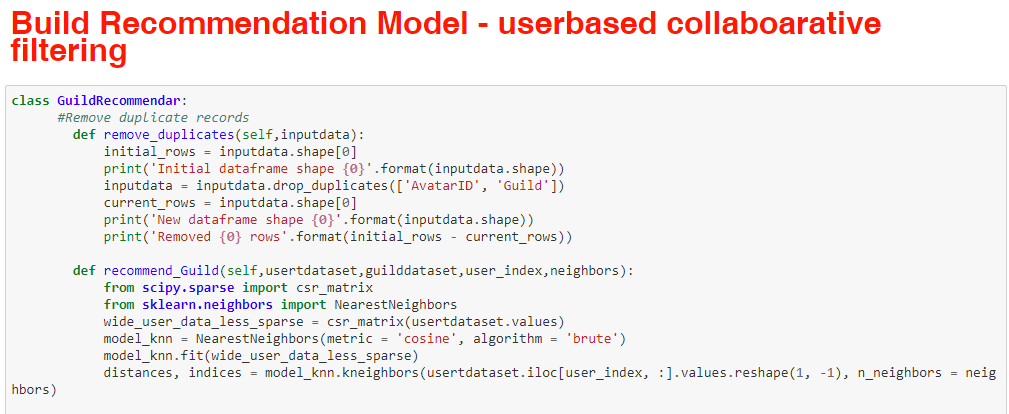


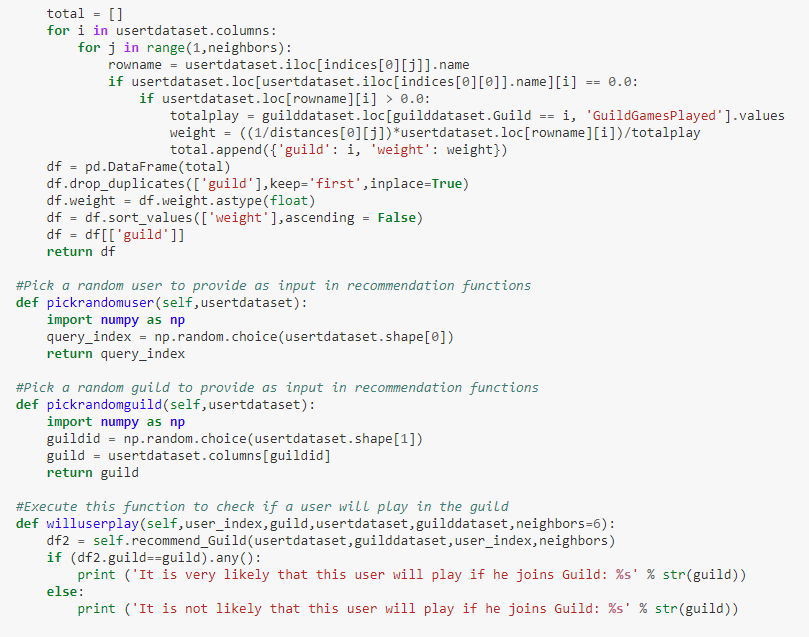






### 5.8 Recommendation Engine – KNN and User Based Collaborative Engine – Feature Set 2







# V. Model Evaluation

## i. Forecast the number of players expected in future time point



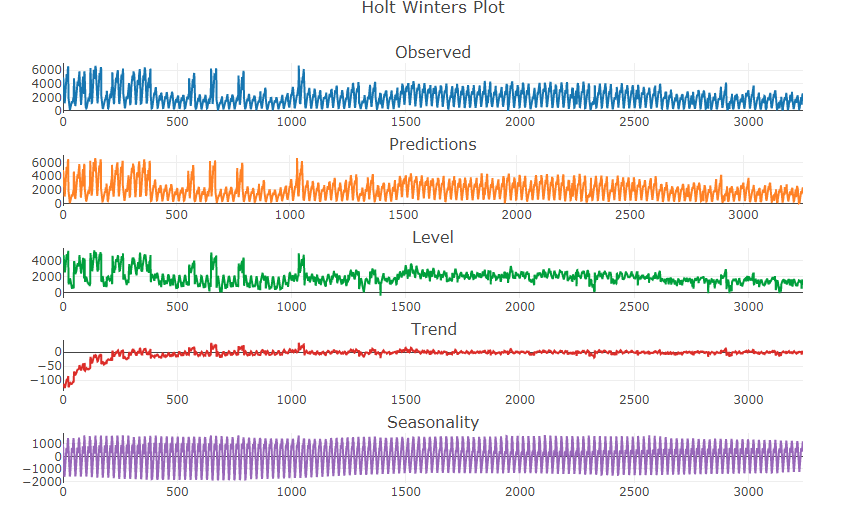


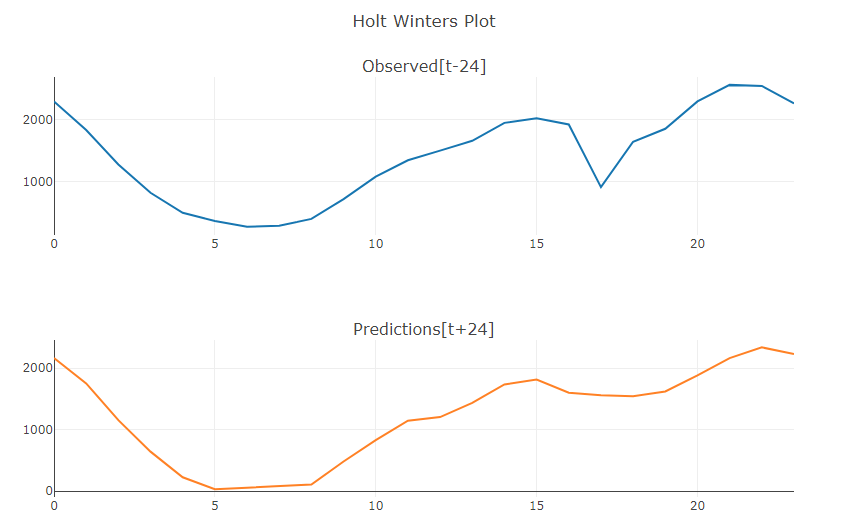
This RMSE and MAPE values are much lesser compared to the values from ARIMA model and LSTM Tensorflow attempted.

Hence Holt Winters Triple Exponential Smoothing - Additive method seems to be more dependable for wow dataset time series forecasting.

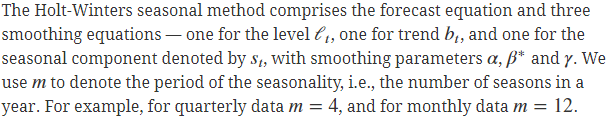
Accuracy of this model is highly dependent on the choice of values chosen for alpha, beta and gamma parameters. Even though there is a function available to calculate this value, choosing the optimal parameters is more based on trial and error.

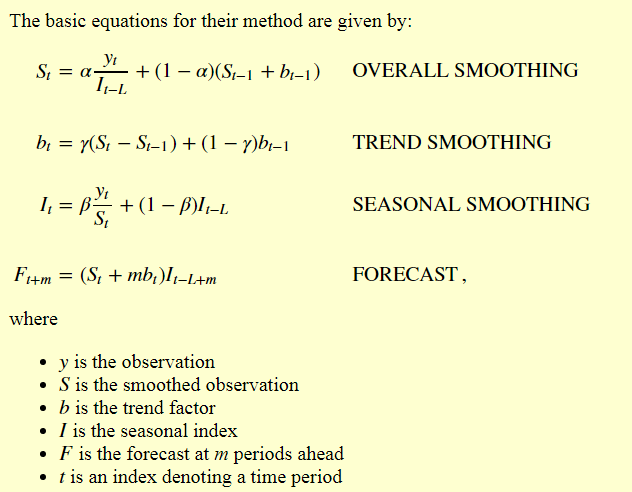
Also, this model is more suitable for forecasting short term time points in future. If the model is attempted to predict for longer term periods in future, accuracy would be very less and values might just show a linear trend since historical data for these kind of games would become stale as the time increases.





### Holt-Winters seasonal method



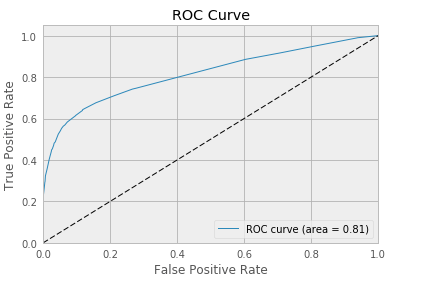


Source: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm

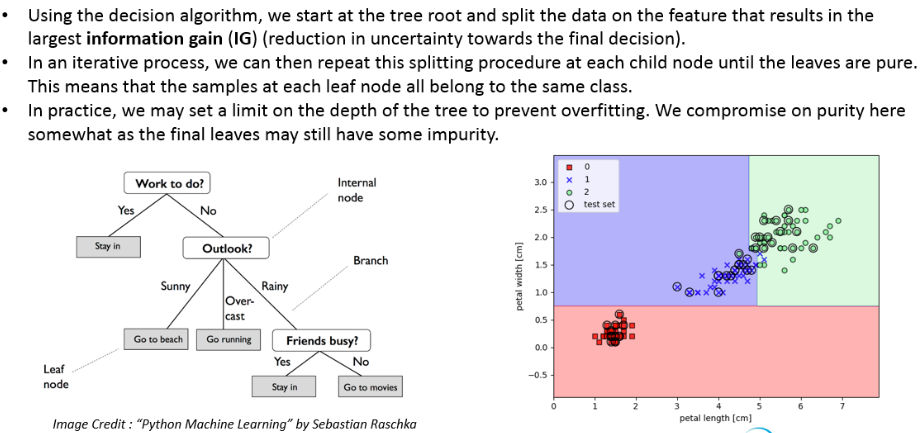
## ii. Predict player churning







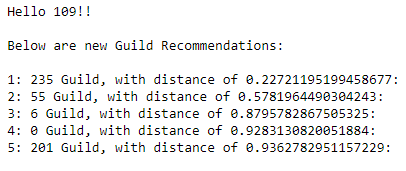
### Decision Tree Classifier

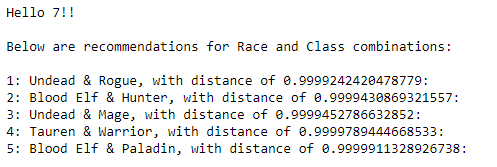


## iii. Recommend Guild, Race/Class to players



Output:





### KNN and User based collaborative filtering

Example:

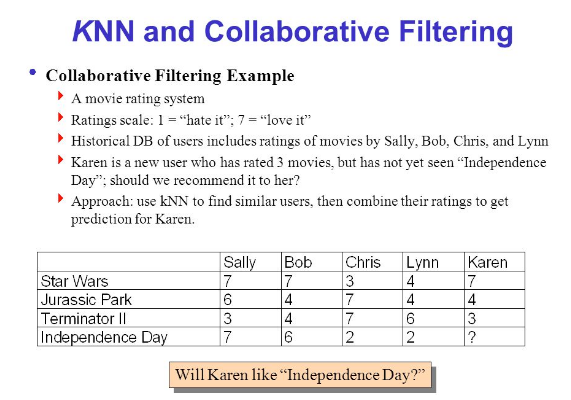


Image credit: http://slideplayer.com/slide/4462580/

# VI. Comparison to benchmark

Models attempted initially were initial benchmarks considered:

[Explained under the section: Appendix -> [Models attempted before](#_Models_attempted_before)]

1. Neural Networks for players behavior pattern recognition
2. ARIMA for time series forecasting

Both the models did not provide expected results or accuracy.

Comparison of Final solution to initial benchmark



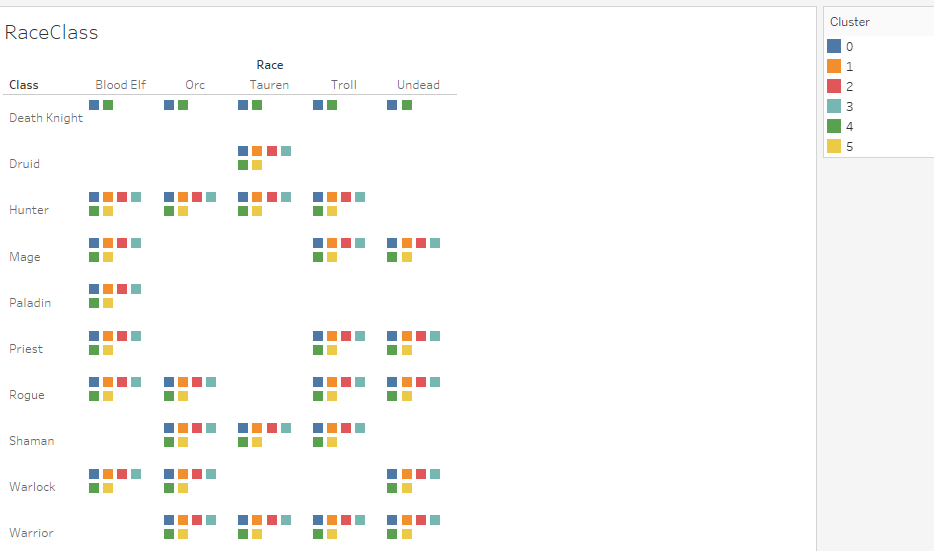
# VII. Visualizations

We have used Plotly and Tableau to perform visualizations on the dataset.

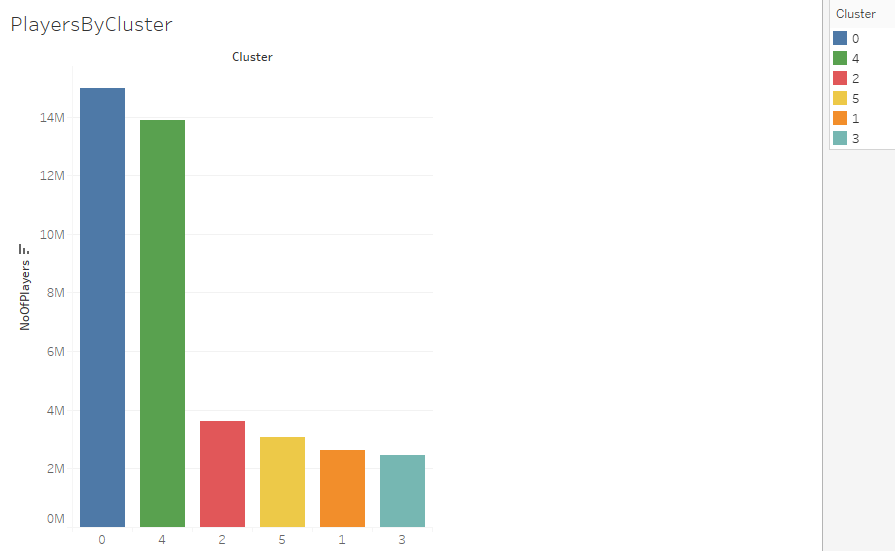
Exploratory visualization on the complete dataset is already provided under the [above section](#_Step_3:_Perform).

Apart from the exploratory visualization of the input dataset, we have also performed visualizations on the intermediate datasets which helped us in proceeding with further model decisions.

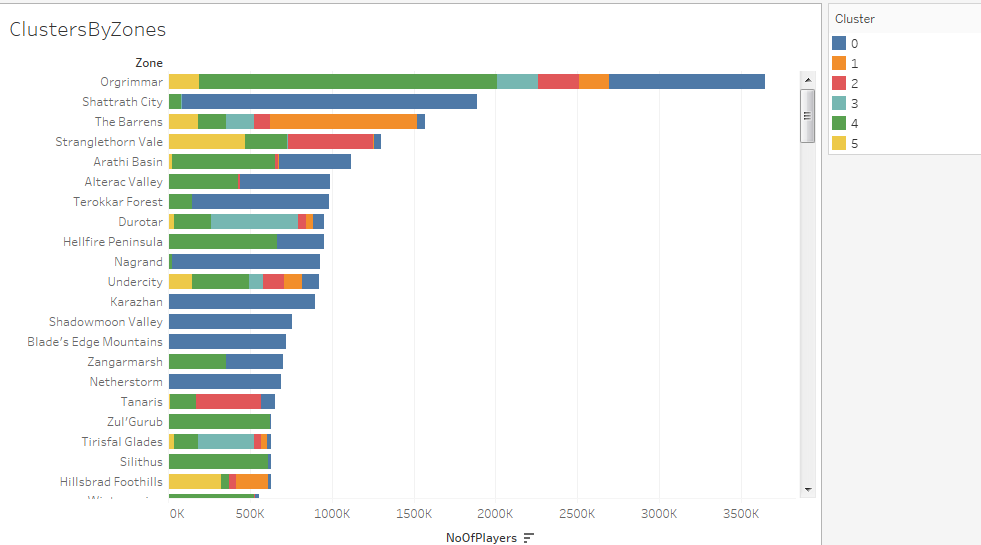
## Cluster Data Analysis



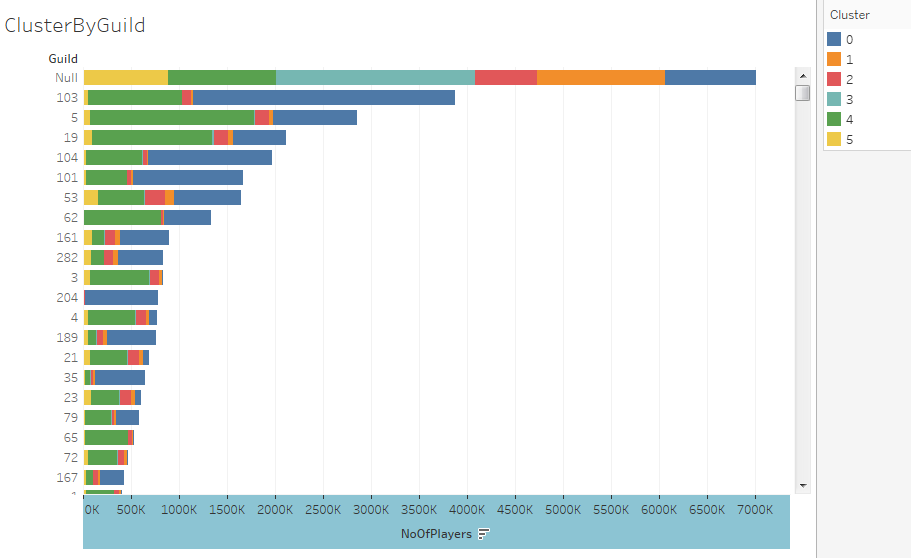
Warrior and Hunter Classes with a combination of BloodElf, Orc and Tauren races are played by maximum players



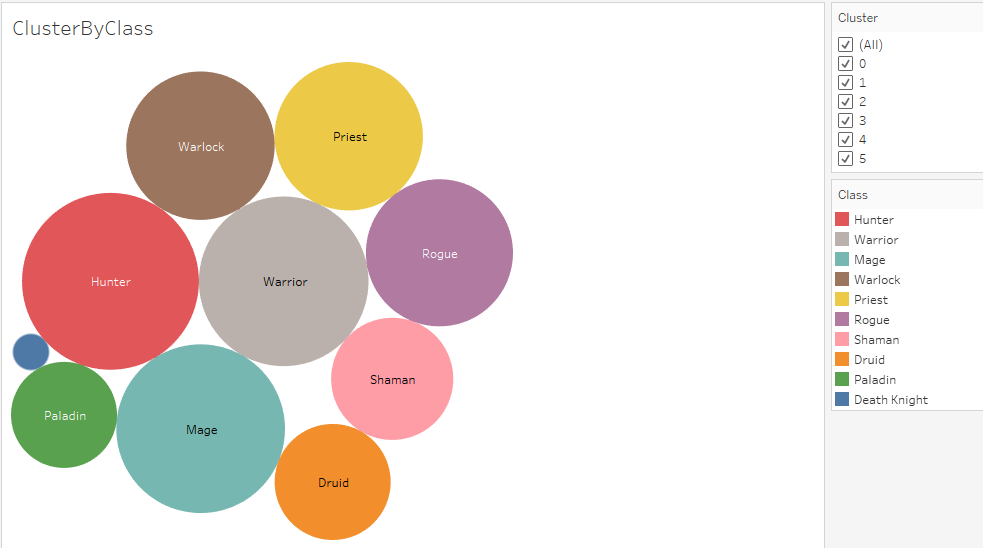
Majority of the players are grouped under Cluster 0

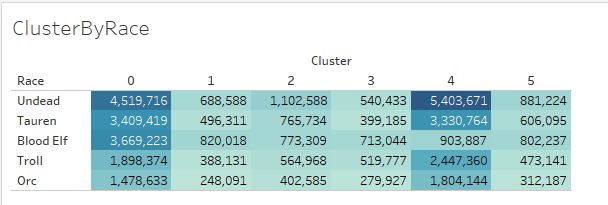


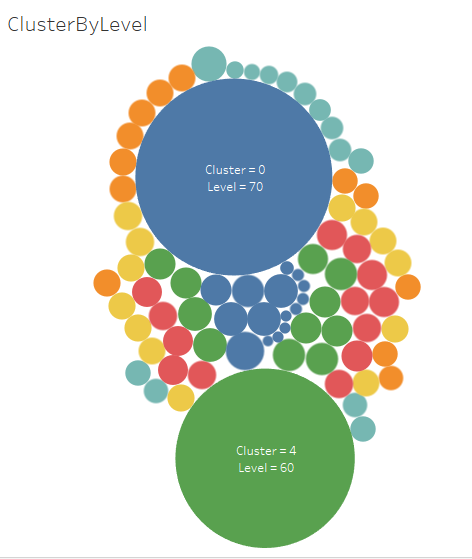
Origrimmar is the most played zone.



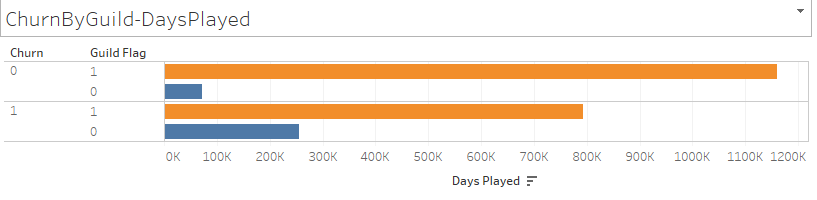
Maximum players played without Guild are sorted into Cluster 3

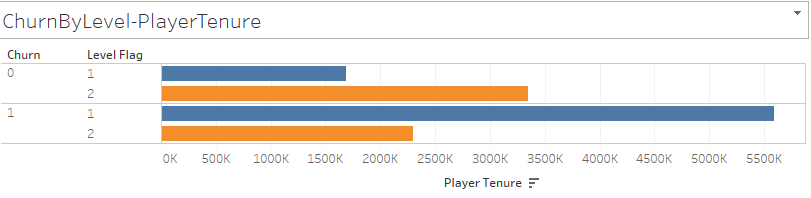


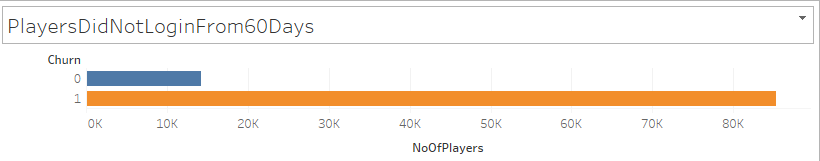


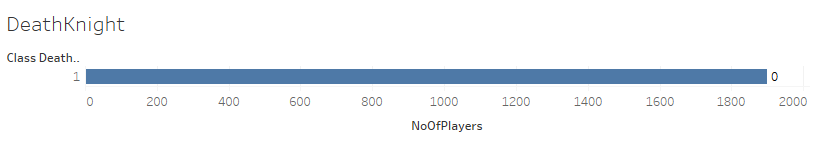


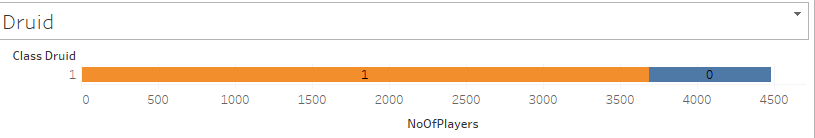
## Churn Data Analysis

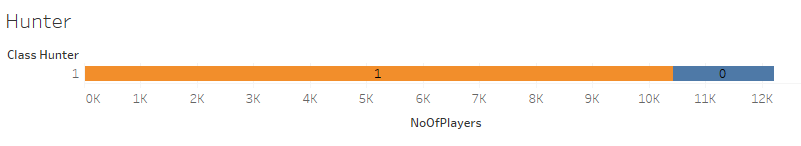
****

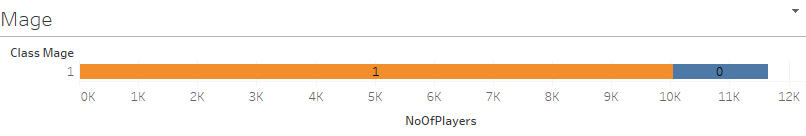
****

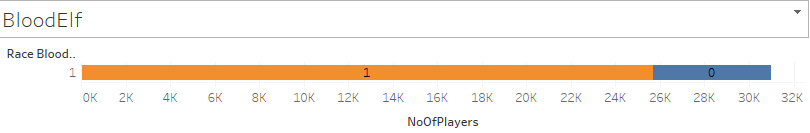


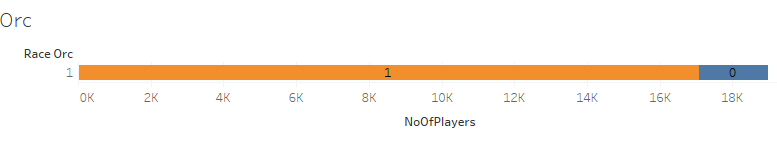












**Insights**

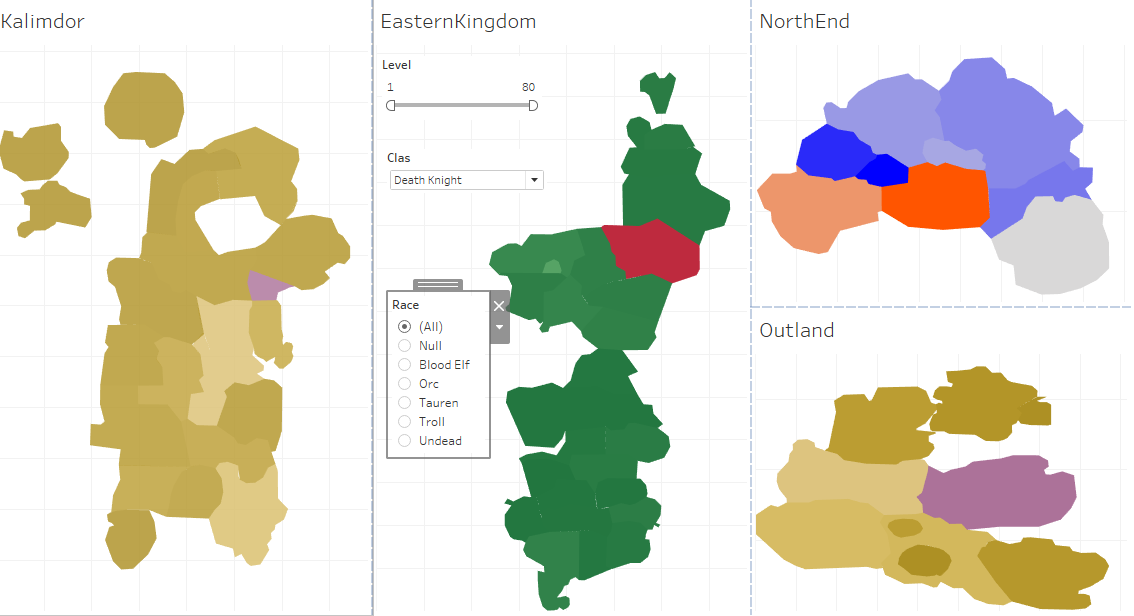
1. Churn is high when players were unable to go to next levels quickly.

Recommendation to Developers: Complexity of the initial levels needs to be brought down to let players play with more interest.

1. Interestingly Death Knight is the Class played by less number of Players but none of the players playing Death Knight have churned.

Recommendation to Developers: Recommend the Class “Death Knight” to more users so that they can explore and play more

## WOW Zone Heat map



# VIII. Implications



Recommendations to Business are provided at a 95% confidence level

# IX. Limitations

1. These models are derived based on the data available between the three year periods of 2006 to 2009.
2. In real world scenario: more recent data would be required to redesign this model and to maintain continuously.
3. Also, the features available in the dataset is very less (7 Features) which makes it challenging to provide more in depth insights and to identify more business problems and solutions.
4. To enhance the solutions provided by us, it would be very essential to capture logs with more information in future.

# X. Closing Reflections

**Learnings:**

1. **Extensive usage of Tableau to bring more meaningful insights from World of Warcraft Logs**
2. **Multiple Time series Forecasting models and their application on WoW logs**
3. **Methods of application of Feature Engineering on WOW dataset since the available features were not enough of successful model building**
4. **Appropriate usage of Model Tuning and Regularization on WoW dataset**

**Things to do differently next time:**

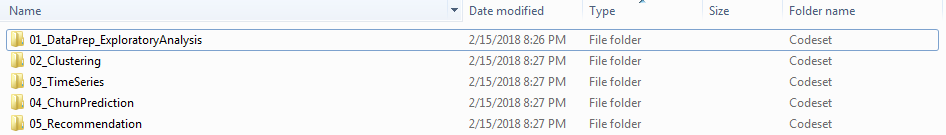
1. **Collect data that has captured more features in future logs to bring more insights**
2. **Explore the usage of Spark R on this dataset**

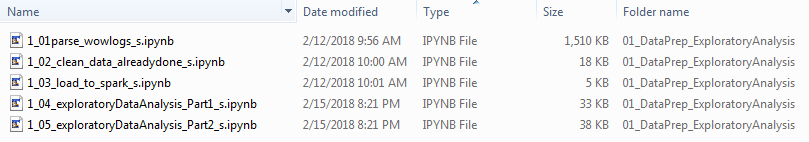
# XI. Code Information

**Input Data**

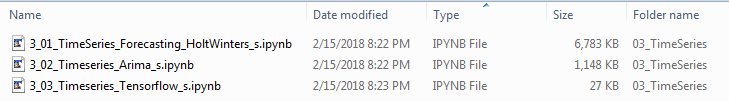
<http://mmnet.iis.sinica.edu.tw/dl/wowah/wowah.rar>

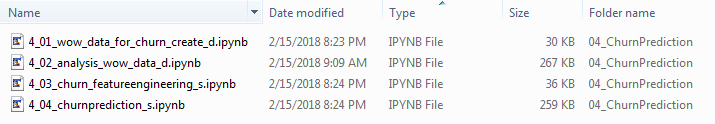
**Jupyter Notebook**

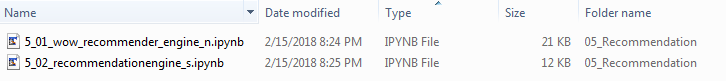




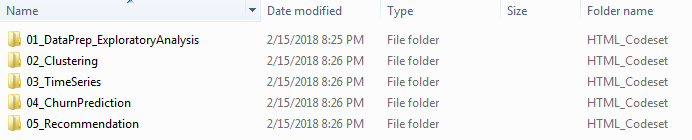




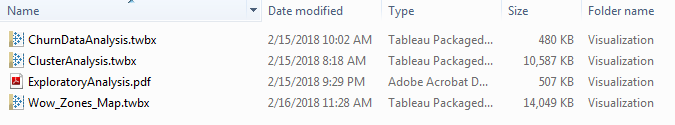




**HTML Files**

****

**Tableau Visualizations**

****

**Project in Github with all of the above code:**

<https://github.com/capstoneaiscientists/Team3GLBDA>

# Appendix

## Other models attempted before interim report

#### 1. Neural networks for pattern recognition - Exploratory

In the process of solving problem statements, we first came up with identifying the game play patterns of players from the dataset.

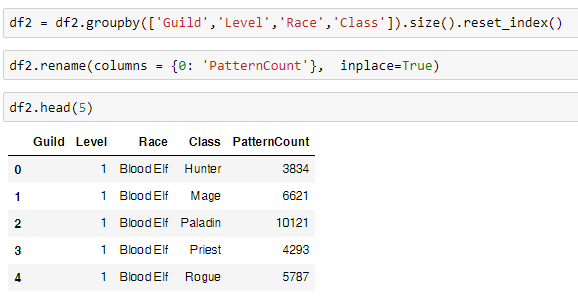
For this process, we have converted the dataset into numerical data that can be provided as input to Pattern recognizing neural networks.

This model would still be part of exploratory data analysis.

Step 1: Read data into a dataframe



Step 2: Group records by Guild, Level, Race and Class and count the number of records following each unique combination



Step 3: Convert string values on the above columns into their encoded numeric value

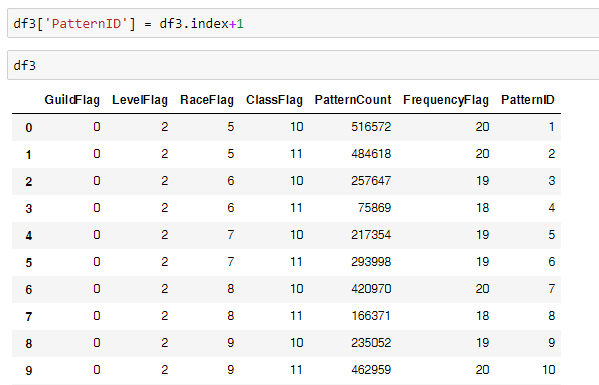




Step 4: Create Frequency Flag by creating a rule for each frequency



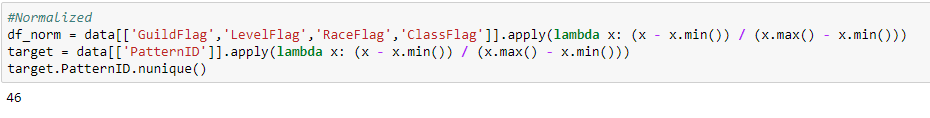
Step 5: Create Pattern ID for each unique pattern

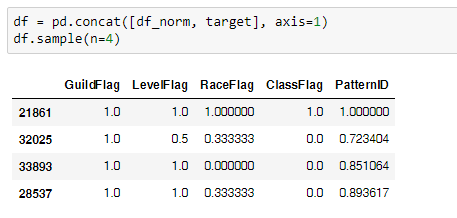


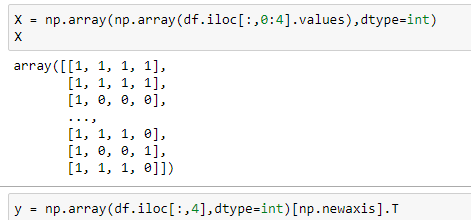
Step 6: Identify the number of unique patterns in the dataset



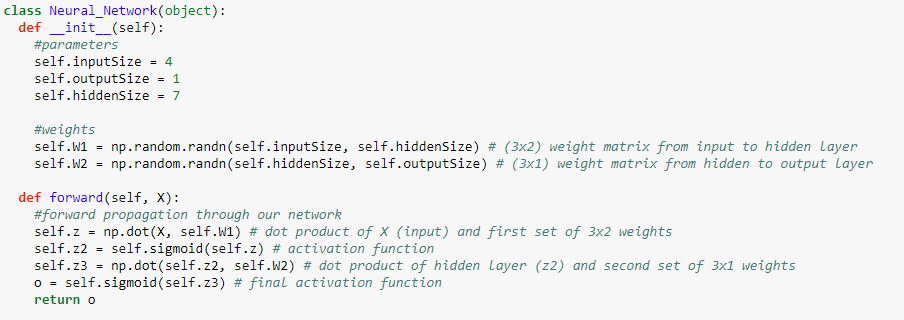
Step 7: Normalize the data and create Train & Test data



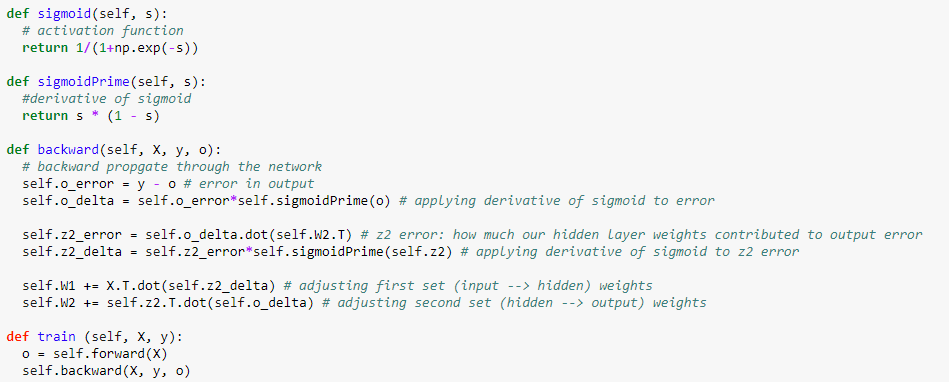




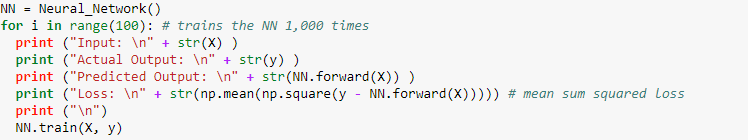
Step 8: Provide inputs to Multiple Back Propagation neural networks and predict patterns (Input Layer = 4, Hidden Layer = 7 and Output Layer = 1)



Step 9: Use sigmoid function to activate the neurons



Step 10: Predict Pattern id and Loss on Test Data



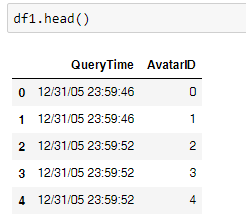


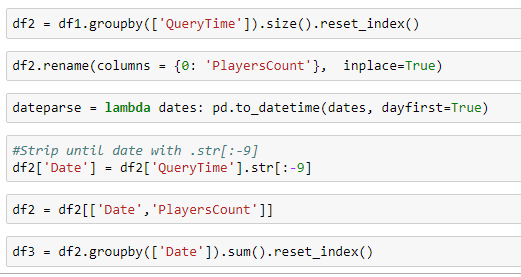
The results of the tests with the neural networks shows how it can be trained to recognize groups of players by their characteristics and learn the unique patterns of Guild, Level, Race, Class combinations of these groups in the game.

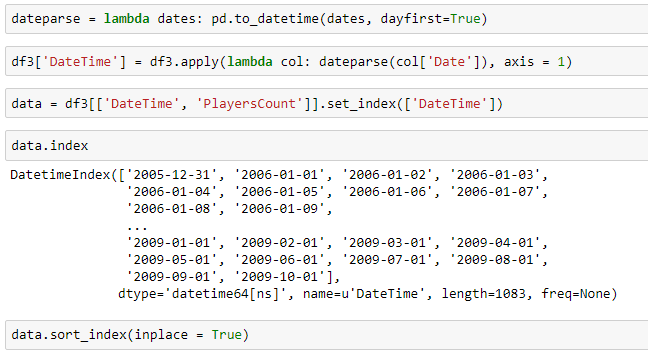
#### 2. Time series forecasting to predict number of players

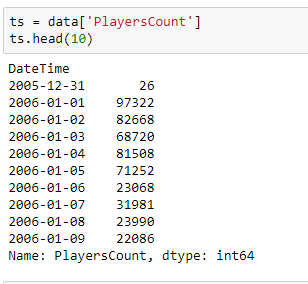
Another algorithm applied so far on this dataset is ARIMA model **Auto-Regressive Integrated Moving Averages** to predict number of players expected on a future time series

Step1: Load and handle Time series data in Python

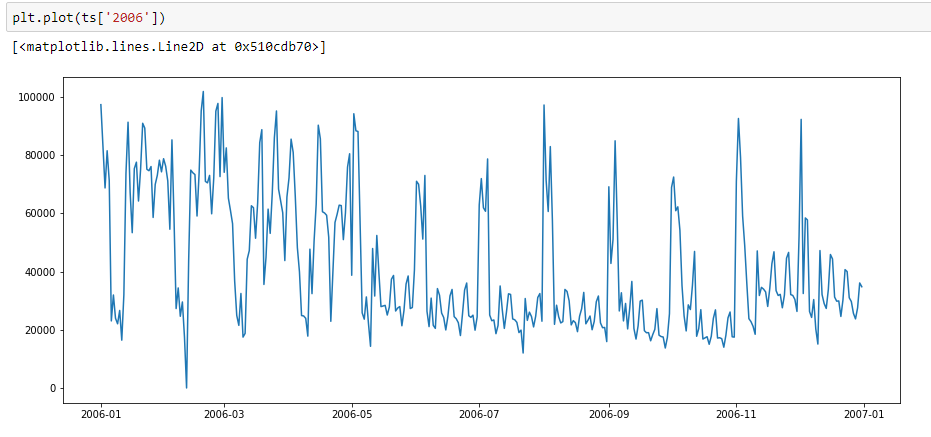




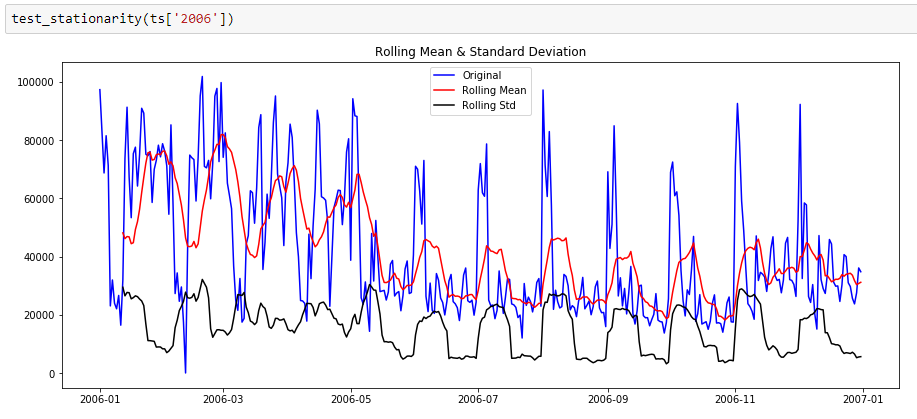




Step 2: Check stationarity of a Time series

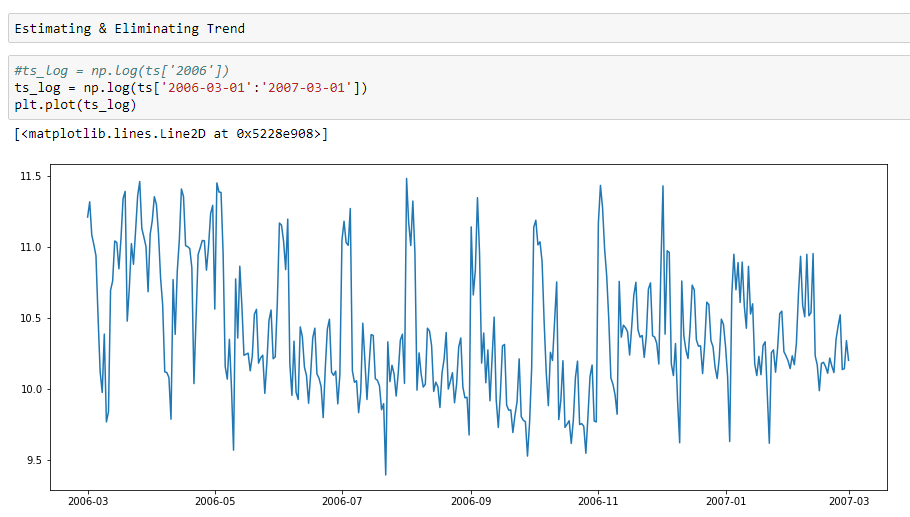


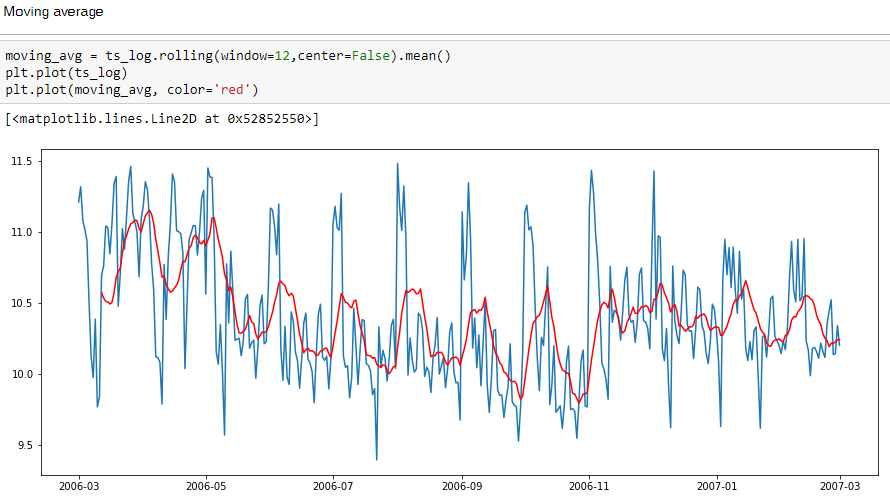


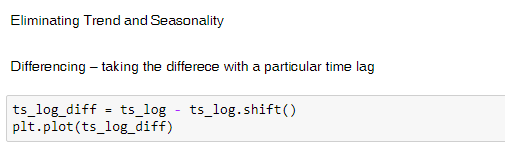


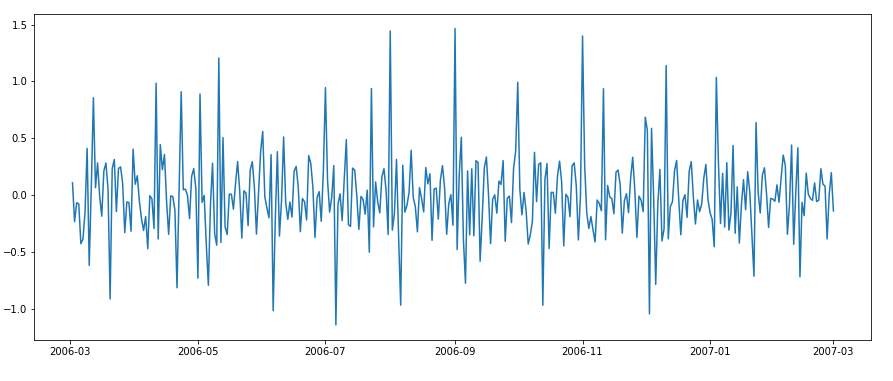


Step 3: Make time series stationary

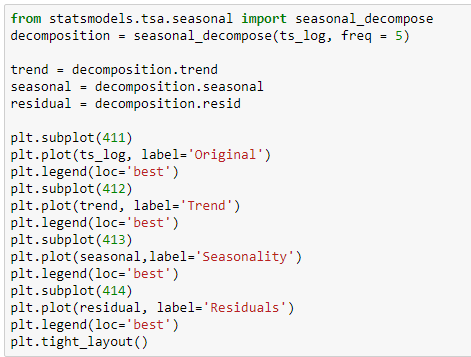


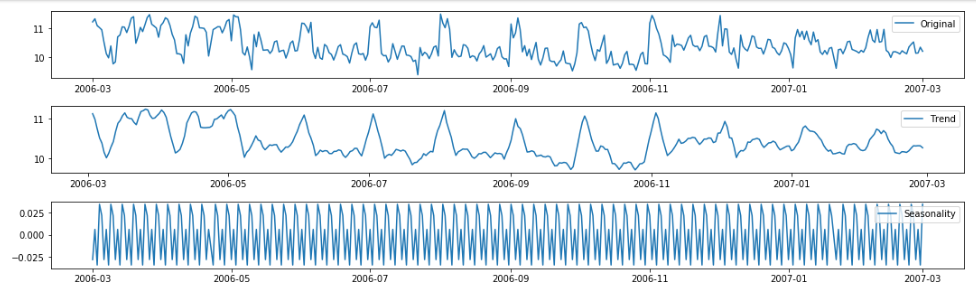


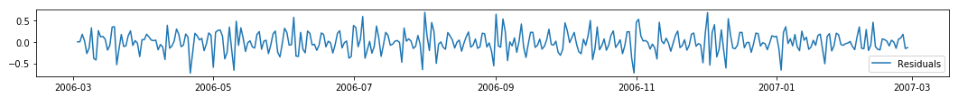






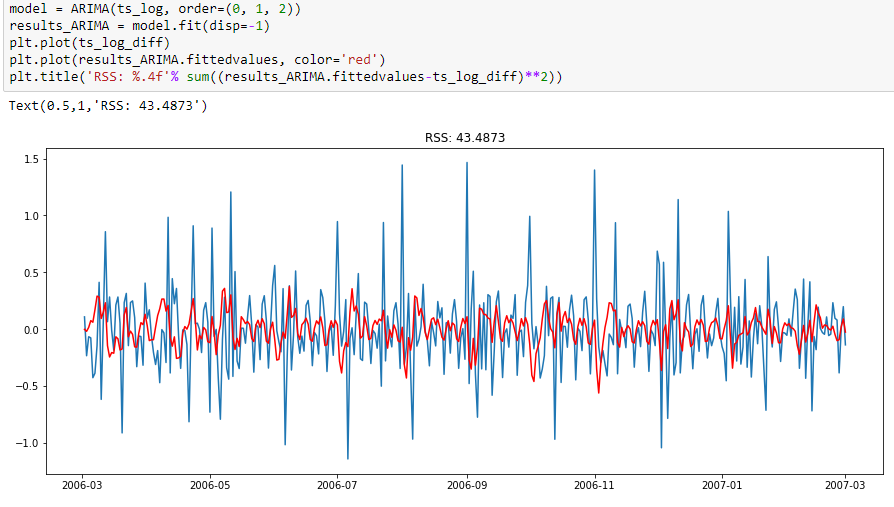


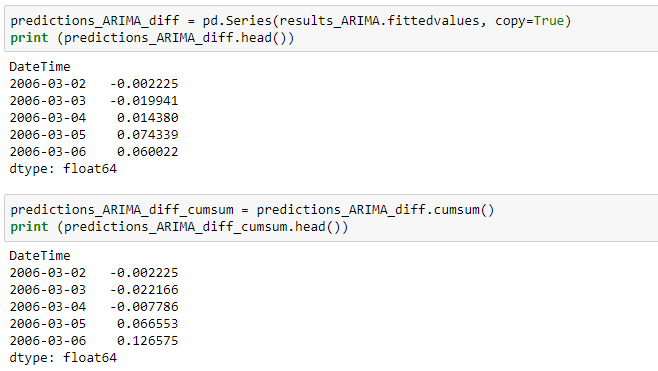


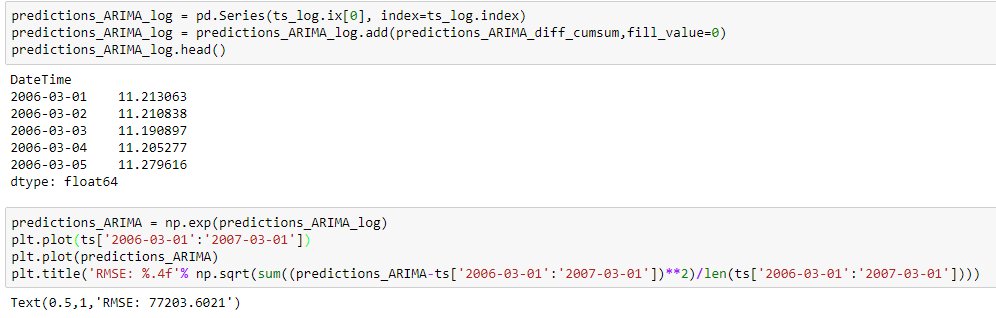


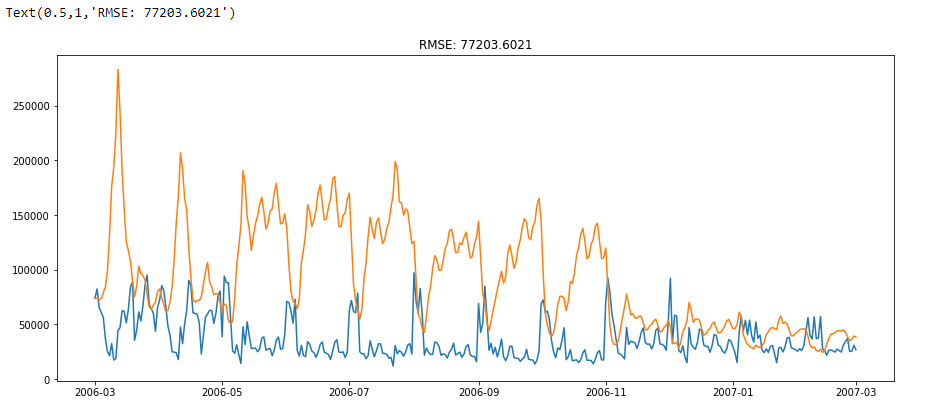
Step 4: Forecast on Time series

Applying ARIMA model









## References

1. **Big Data Analytics Using Neural networks**

***Author****: Chetan Sharma*

***Master’s Thesis:*** *San Jose State University*

1. **Game Analytics - Maximizing the Value of Player Data**

***Authors:*** *Magy Seif El-Nasr, Anders Drachen and Alessandro Canossa*

***Publisher:*** *Springer*

1. **Big Data Analytics in Cloud Gaming**

***Authors:*** *Victor Perazzolo Barros and Pollyana Notargiacomo*

**Paper:** 2016 IEEE International Conference on Big Data

1. **Setting Players’ Behaviors in World of Warcraft through Semi-Supervised Learning*****Authors:*** *Marcelo Souza Nery, Victor do Nascimento Silva, Roque Anderson S. Teixeira and Adriano Alonso Veloso*