BEHAVIORAL PATTERN RECOGNITION OF MULTIPLAYER ONLINE ROLEPLAYING GAME PLAYERS USING BIG DATA ANALYTICS AND MACHINE LEARNING

Team members:

Sulekha AloorRavi

Niharika Thanavarapu

Deepa Venugopal

Lakshmipriya M

Mentor:

Dr. Narayana D

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.
- Great volumes of data are generated all the time in gaming 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

Problem Statement

- PERFORM EXPLORATORY ANALYSIS
  - To cluster players into different groups based on features in dataset
  - To analyze and visualize timeline patterns of players by different groups and parameters
  - To create heat map based on the gaming zones
  - To visualize patterns based on player Guilds

- PERFORM PREDICTIVE ANALYTICS BY APPLYING MACHINE LEARNING MODELS
  - Forecast the number of players expected in future time point to plan resource capacity
  - Predict player churning to come up with steps to avoid future churn
  - Recommend guilds [groups to join] to players for effective gaming and to minimize churn

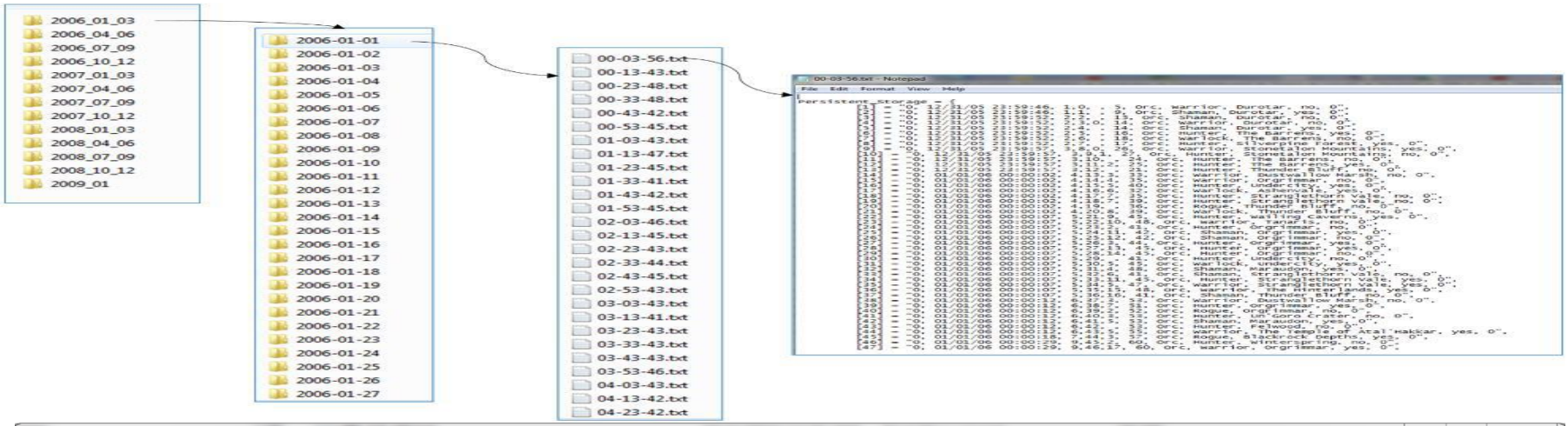
Data

- 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

Input Dataset



00-03-56.txt - Notepad

File Edit Format View Help

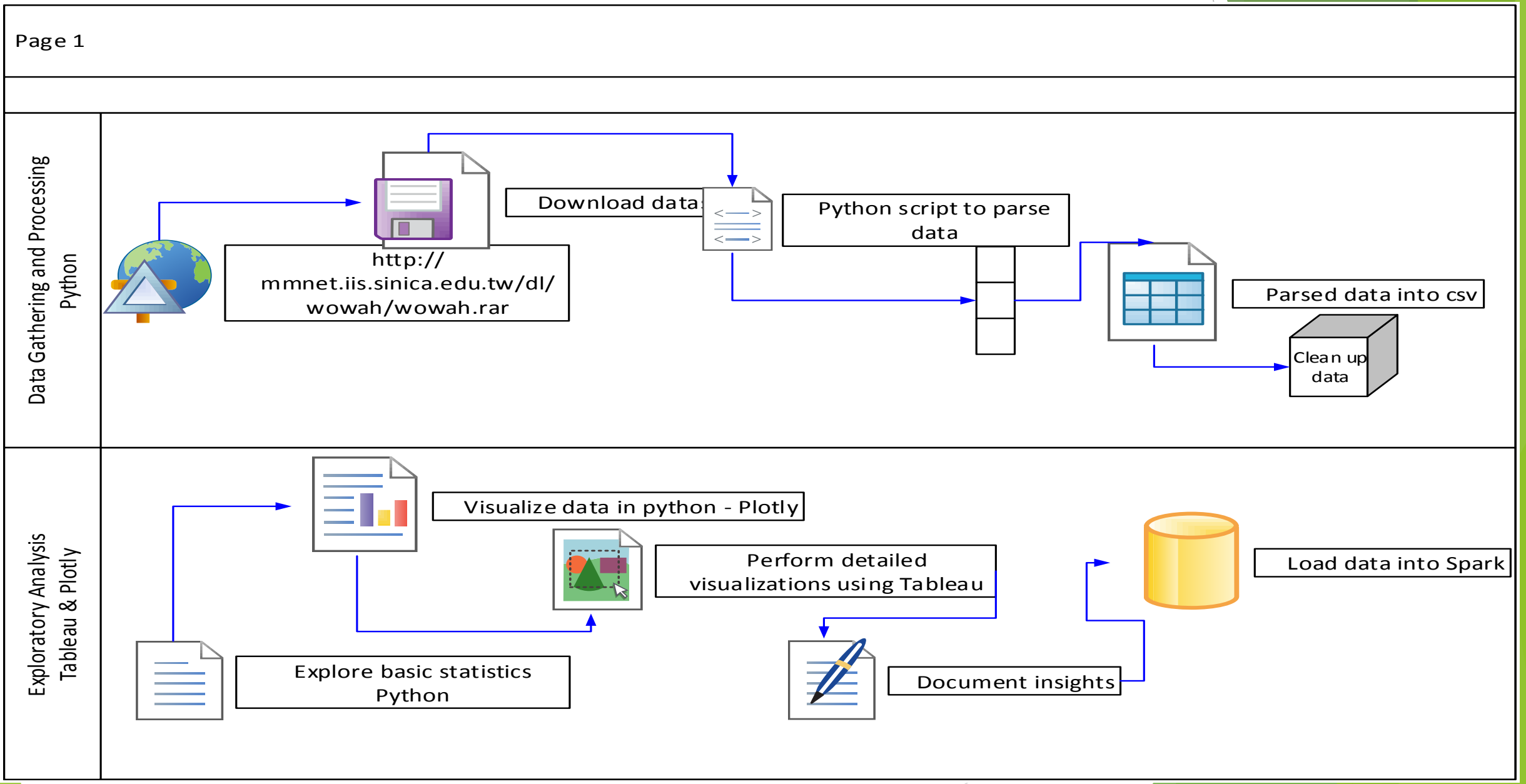
```
Persistent_Storage = {
        [1] = "0, 12/31/05 23:59:46, 1,0, , 5, Orc, Warrior, Durotar, no, 0",
        [2] = "0, 12/31/05 23:59:46, 1,1, , 9, Orc, Shaman, Durotar, yes, 0",
        [3] = "0, 12/31/05 23:59:52, 2,2, , 13, Orc, Shaman, Durotar, no, 0",
        [4] = "0, 12/31/05 23:59:52, 2,3,0, 14, Orc, Warrior, Durotar, no, 0",
        [5] = "0, 12/31/05 23:59:52, 2,4, , 14, Orc, Shaman, Durotar, yes, 0",
        [6] = "0, 12/31/05 23:59:52, 2,5, , 16, Orc, Hunter, The Barrens, yes, 0",
        [7] = "0, 12/31/05 23:59:52, 2,6, , 18, Orc, Warlock, The Barrens, no, 0",
        [8] = "0, 12/31/05 23:59:52, 2,7, , 17, Orc, Hunter, Silverpine Forest, yes, 0",
        [9] = "0, 12/31/05 23:59:57, 3,8,0, 26, Orc, Warrior, Stonetalon Mountains, yes, 0",
        [10] = "0, 12/31/05 23:59:57, 3,9,1, 27, Orc, Hunter, Stonetalon Mountains, no, 0",
```

00-00-16.txt - Notepad

File Edit Format View Help

```
Persistent_Storage = {
        "0, 10/16/08 23:55:52, 1,78433, , 22, Orc, Hunter, Wailing Caverns, HUNTER, 0", -- [1]
        "0, 10/16/08 23:55:52, 1,71698,189, 51, Orc, Warrior, Tanaris, WARRIOR, 0", -- [2]
        "0, 10/16/08 23:55:52, 1,83311, , 1, Orc, Warrior, Durotar, WARRIOR, 0", -- [3]
        "0, 10/16/08 23:55:52, 1,2232,171, 37, Orc, Hunter, Stranglethorn Vale, HUNTER, 0", -- [4]
        "0, 10/16/08 23:55:52, 1,67891,342, 46, Orc, Shaman, Stranglethorn Vale, SHAMAN, 0", -- [5]
        "0, 10/16/08 23:55:52, 1,82547, , 13, Orc, Hunter, The Barrens, HUNTER, 0", -- [6]
        "0, 10/16/08 23:55:57, 2,74746,4, 62, Orc, Hunter, Silithus, HUNTER, 0", -- [7]
        "0, 10/16/08 23:55:57, 2,27045,115, 61, Orc, Hunter, The Barrens, HUNTER, 0", -- [8]
        "0, 10/16/08 23:56:02, 3,4876,245, 68, Orc, Hunter, Hall of Legends, HUNTER, 0", -- [9]
        "0, 10/16/08 23:56:02, 3,65943,273, 69, Orc, Hunter, Shattrath City, HUNTER, 0", -- [10]
```

Overview of the Final Process

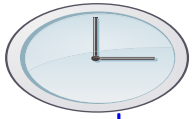


Overview of the Final Process...

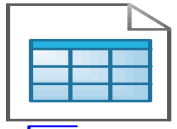
Page 2

Feature Engineering Spark

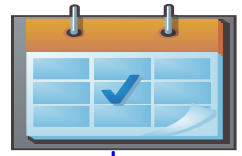
Load data into Spark



Modify data into time series suitable for forecasting

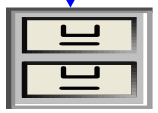


Derive new features to do Churn prediction

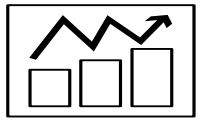


Create Sparse Matrix for recommendation engine

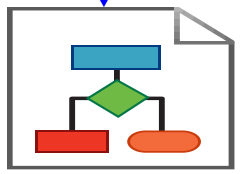
Model selection, Tuning and Evaluation Spark ML



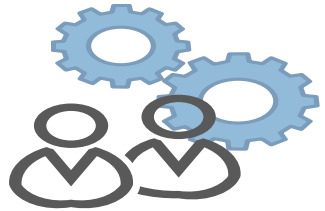
Kmeans Clustering



1. Holt Winters
2. Arima
3. LSTM Tensowflow



1. Decision Tree Classifier
2. Random Forest Classifier
3. Gradient Boosted Tree Classifier



1. K-Nearest Neighbors
2. User based collaborative Filtering

Final Reporting

Prepare final report with:

1. pyspark and python programs
2. Code input/output HTMLs
3. Input Datasets
4. Tableau visualizations
5. Final PDF Report with business insights

Log Parsing

```
parse = wow_parser()
```

```
dirpath = "H:\WoWAH"
outputpath = "H:\Output\wowlogs.csv"
parse.parse_logs(root_dir = dirpath, output_file = outputpath)
```

	QueryTime	QuerySeq	AvatarID	Guild	Level	Race	Class	Zone
0	12/31/05 23:59:46	1	0		5	Orc	Warrior	Durotar
1	12/31/05 23:59:46	1	1		9	Orc	Shaman	Durotar
2	12/31/05 23:59:52	2	2		13	Orc	Shaman	Durotar
3	12/31/05 23:59:52	2	3	0	14	Orc	Warrior	Durotar
4	12/31/05 23:59:52	2	4		14	Orc	Shaman	Durotar

Data Cleanup

Following values are incorrect Warcraft races:
'373族', '547人', '3033','27410', '74622妖'
Let us look at the records which have these incorrect races.

```
df_incorrect_race = df[df['Race'].isin(['373族', '547人', '3033','27410', '74622妖'])]
```

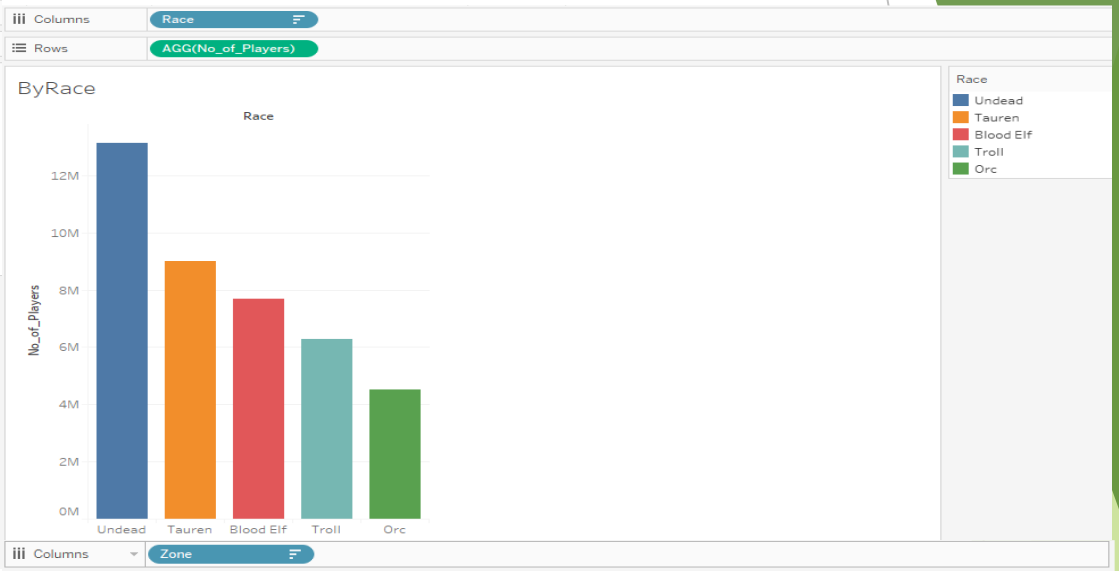
```
df_incorrect_race.AvatarID.unique()
```

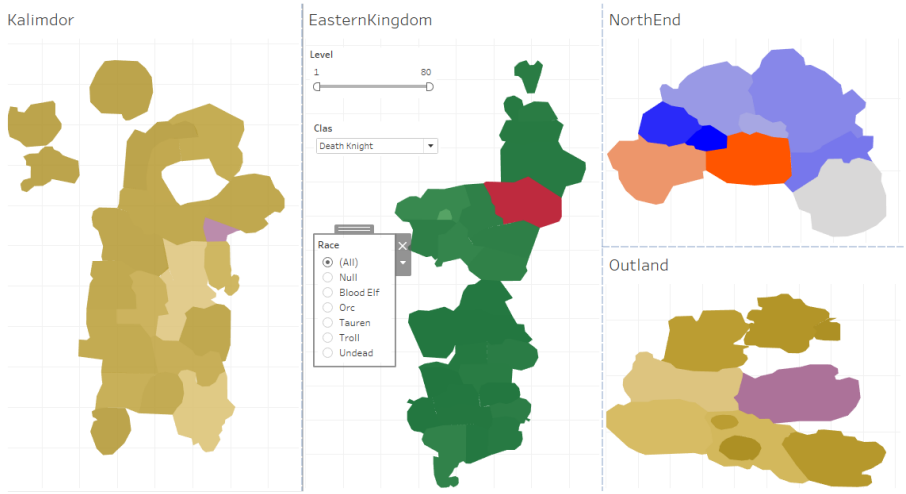
```
array([  373,   547,  3033, 27410, 74622], dtype=int64)
```

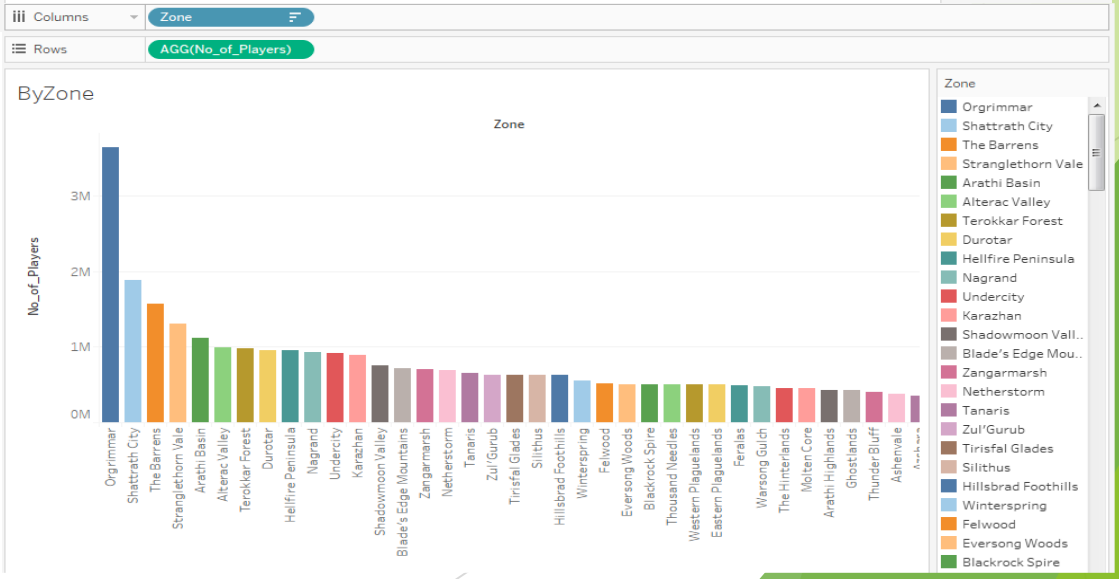
```
df_incorrect_race.count()
```

```
QueryTime    50085
QuerySeq     50085
AvatarID     50085
Guild        50085
Level        50085
Race         50085
Class        50085
Zone         50085
dtype: int64
```

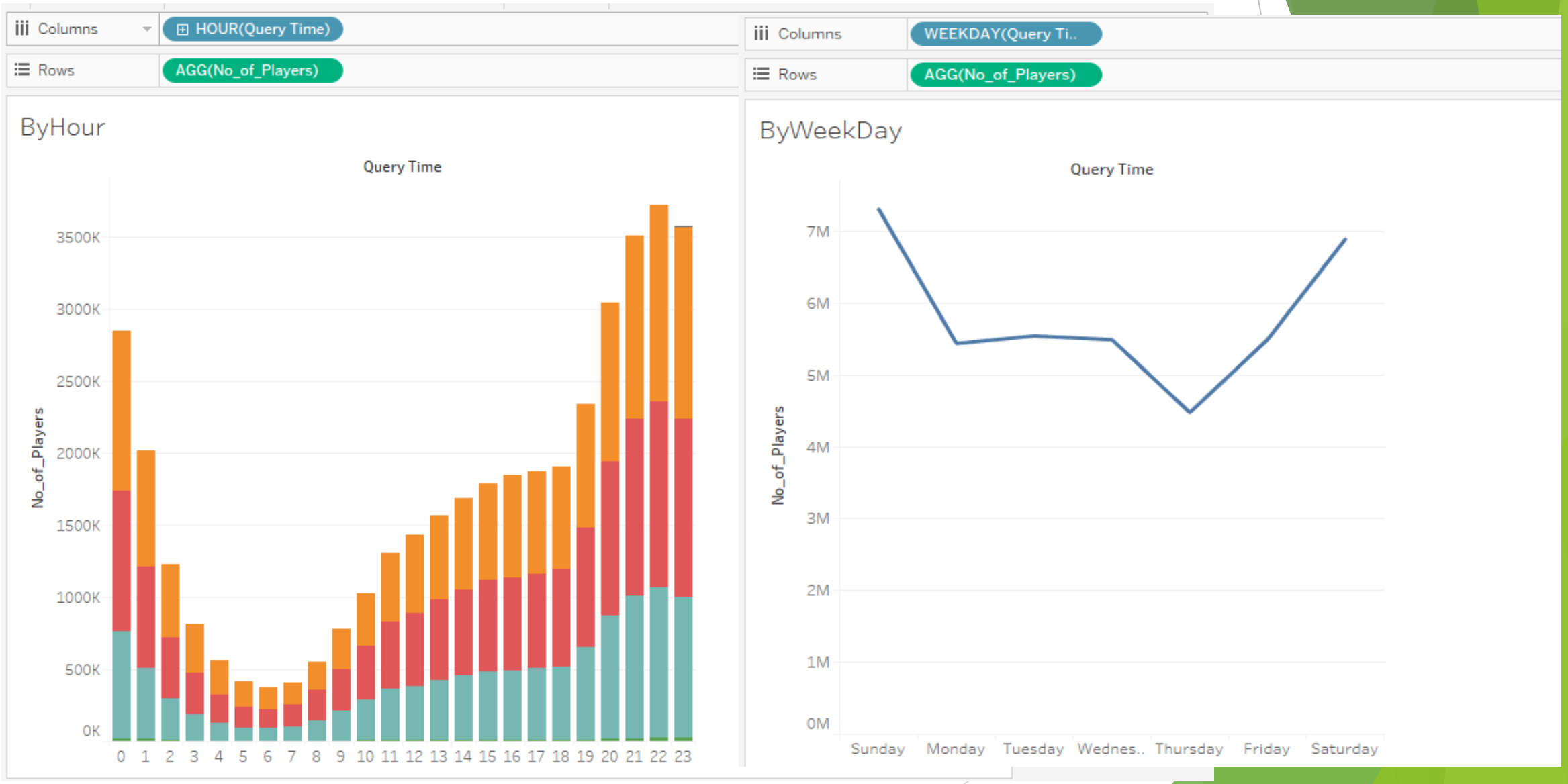
Exploratory Analysis







Exploratory Analysis…



Feature Engineering

- New Features are created from existing features to make the data suitable for:
  - Time series Forecasting
  - Player Churn prediction
  - Recommendation Engine

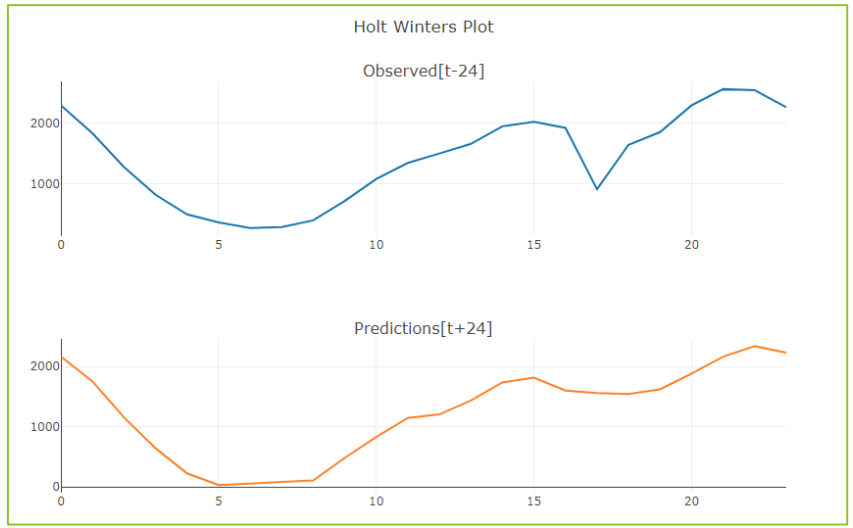
```
FinalChurnDF
```

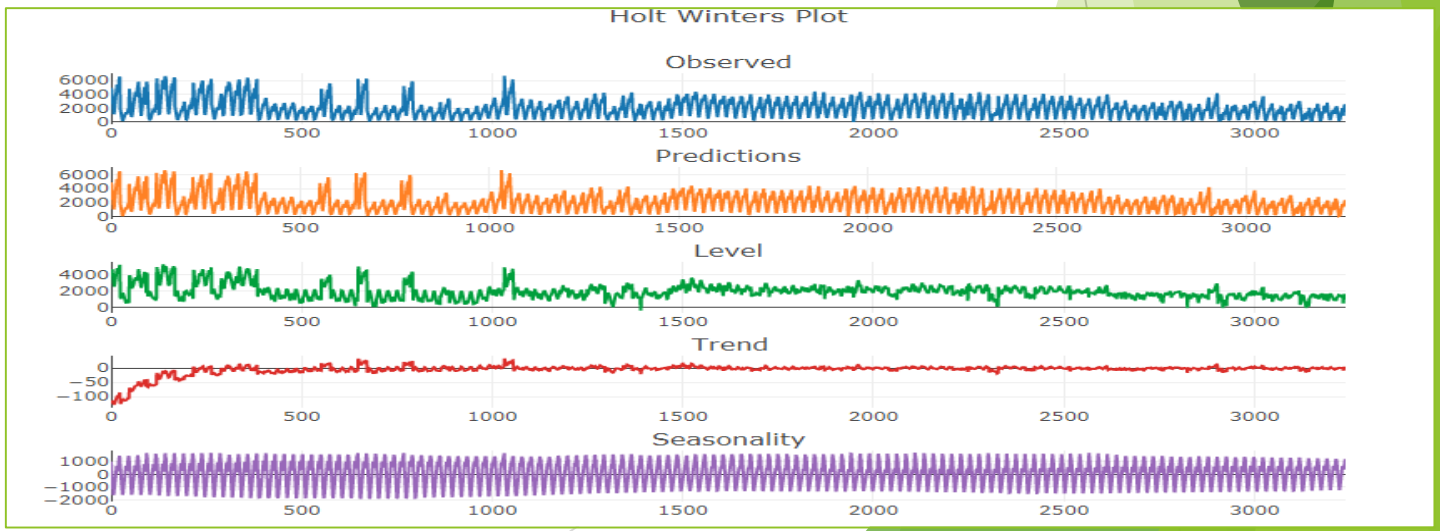
```
DataFrame[AvatarID: int, Class_Warlock: int, Class_Druid: int, Class_Hunter: int, Class_DeathKnight: int, Class_Paladin: int,
lass_Rogue: int, Class_Mage: int, Class_Priest: int, Class_Warrior: int, Class_Shaman: int, Race_Orc: int, Race_Tauren: int, R
ce_Undead: int, Race_BloodElf: int, Race_Troll: int, ZonesPlayed: bigint, LevelFlag: int, GuildFlag: int, DaysPlayed: bigint,
layerTenure: double, Churn: int]
```

Model building and Evaluation

- Time series Forecasting

Problem Statement	Evalation parameter	Models Attempted	Best Score	Final Model	Code
2.1 Forecast the number of players expected in future time point	MAPE $\left(\frac{1}{n}\sum\frac{\lvert Actual - Forecast\rvert}{\lvert Actual\rvert}\right)*100$ RMSE $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$	Holt Winters ARIMA LSTM Tensowflow	Forecasting MAPE: 0.65 RMSE: 1087.3 Future 24 Periods Prediction MAPE: 51.5 RMSE: 2974.26	Holt Winters	3_01_TimeSeries_Forecasting_HoltWinters_s.ipynb



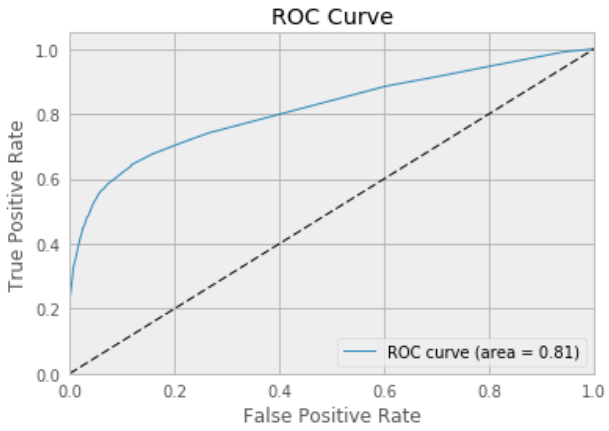


Model building and Evaluation…

- Churn prediction

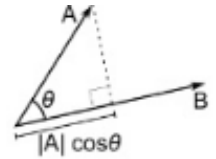
Problem Statement	Evalation parameter	Models Attempted	Best Score	Final Model	Code
2.2 Predict player churning	$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$ $Precision = \frac{T_p}{T_p + F_p}$ $Recall = \frac{T_p}{T_p + T_n}$	Decistion Tree Classifier Random Forest Classifier Gradient-boosted tree classifier	Accuracy: 0.81 Precision: 0.88 Recall: 0.57 +----------+----+----+ |Churn_prediction| 0.0| 1.0| +----------+----+----+ | 1| 488|3712| | 0|2763|1491| +----------+----+----+	Decistion Tree Classifier	Featureset1 - 4_02_analysis_wow_data_d.ipynb Featureset2- 4_04_churnpredictio n_s.ipynb

Note:
Have performed an extensive feature engineering and created 22 new features from existing 7 features available in the dataset.



Model building and Evaluation…

- Recommendation Engine

Models Applied	Calculation parameter	Code
K-Nearest Neighbors User based collaborative filtering	$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$ 	5_01_wow_recommender_engine_n.ipynb 5_02_recommendationengine_s.ipynb

```
Hello 109!!

Below are new Guild Recommendations:

1: 235 Guild, with distance of 0.22721195199458677:
2: 55 Guild, with distance of 0.5781964490304243:
3: 6 Guild, with distance of 0.8795782867505325:
4: 0 Guild, with distance of 0.9283130820051884:
5: 201 Guild, with distance of 0.9362782951157229:
```

Clustering of data for visualization

KMeans clustering

```
import pyspark.ml.clustering as clus
kmeans = clus.KMeans(k = 6,featuresCol='features')
stages += [kmeans]
```

```
# Create a Pipeline.
pipeline = Pipeline(stages=stages)
```

```
pipelineModel = pipeline.fit(cluster_df)
```

```
cluster_df = pipelineModel.transform(cluster_df)
```

```
cluster_df.printSchema()
```

```
root
 |-- Guild: string (nullable = true)
 |-- Level: integer (nullable = true)
 |-- Race: string (nullable = true)
 |-- Class: string (nullable = true)
 |-- AvatarID: integer (nullable = true)
 |-- Zone: string (nullable = true)
 |-- GuildIndex: double (nullable = true)
 |-- GuildclassVec: vector (nullable = true)
 |-- RaceIndex: double (nullable = true)
 |-- RaceclassVec: vector (nullable = true)
 |-- ClassIndex: double (nullable = true)
 |-- ClassclassVec: vector (nullable = true)
 |-- features: vector (nullable = true)
 |-- prediction: integer (nullable = true)
```

Class distribution in Cluster 3

Level distribution across Clusters 2 & 3

Recommendations to Business

Insights from Prediction Models and Exploratory Analysis	Recommendations to Business
Hunter is the class chosen by most of the players	Develop classes similar to Hunter
Death Knight is chosen by least number of players but interestingly no player playing Death Knight Churned from 3 years	Recommend Death Knight to more users through Recommendation Engine by adding more weightage to it.
Undead is the Race chosen by most of the players	Recommend Undead combining with less played classes to more users.
Orc is chosen by least number of players. More players playing Orc have Churned	Improvise Orc to make it more interesting
Orgrimmar is the Zone chosen by most of the players	Orgimmar is one of the main cities with very interesting events, replicate similar events in other Zones too.
7,014,160 players are playing without joining any guilds. Players without Guilds have Churned more	Recommend more active Guilds to players
January is the month with most players and October is with least players every year	Plan Server and IT support resource availability to be high during these months
Sunday and Saturday (Weekends) are the days with more players and Thursday is the day with least players every week	Plan Server and IT support resource availability to be high during these days based on the outcome of Holt Winters Model
10:00 PM is the most played hour in a day and 6:00 AM is the least played	Do not plan any maintenance or downtime during 10:00 PM, Plan it around 6:00 AM

Limitations

- These models are derived based on the data available between the three year periods of 2006 to 2009.
- In real world scenario: more recent data would be required to redesign this model and to maintain continuously.
- 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.
- To enhance the solutions provided by us, it would be very essential to capture logs with more information in future.

Closing Reflections

- **Learnings:**
  - Extensive usage of Tableau to bring more meaningful insights from World of Warcraft Logs
  - Multiple Time series Forecasting models and their application on WoW logs
  - Methods of application of Feature Engineering on WOW dataset since the available features were not enough of successful model building
  - Usage of Spark ML for Model building

- **Things to do differently next time:**
  - Collect data that has captured more features in future logs to bring more insights
  - Explore the usage of Spark R for model building and Hive database for Feature Engineering

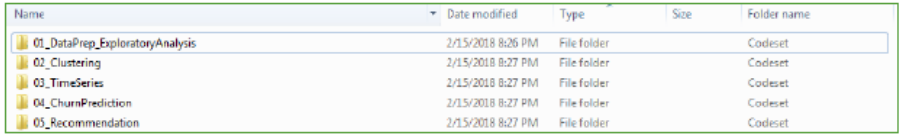
Code Repository

XI. CODE INFORMATION

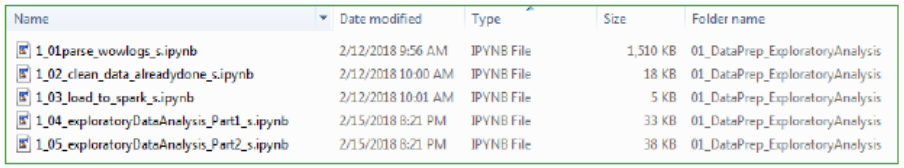
Input Data

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

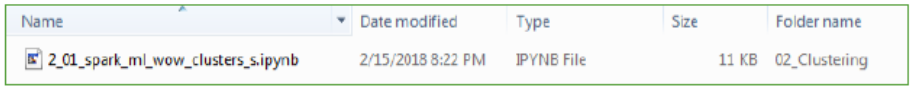
Jupyter Notebook



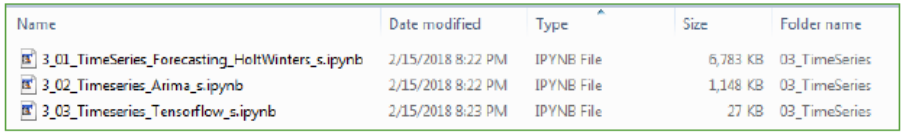
Name	Date modified	Type	Size	Folder name
01_DataPrep_ExploratoryAnalysis	2/15/2018 8:26 PM	File folder		Codeset
02_Clustering	2/15/2018 8:27 PM	File folder		Codeset
03_TimeSeries	2/15/2018 8:27 PM	File folder		Codeset
04_ChurnPrediction	2/15/2018 8:27 PM	File folder		Codeset
05_Recommendation	2/15/2018 8:27 PM	File folder		Codeset



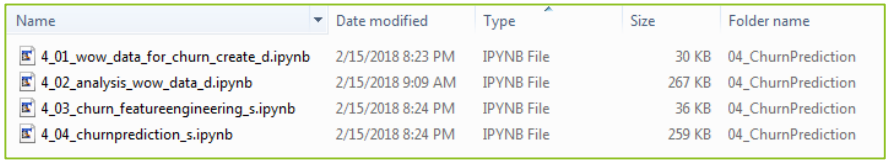
Name	Date modified	Type	Size	Folder name
1_01parse_wowlogs_s.ipynb	2/12/2018 9:56 AM	IPYNB File	1,510 KB	01_DataPrep_ExploratoryAnalysis
1_02_clean_data_alreadydone_s.ipynb	2/12/2018 10:00 AM	IPYNB File	18 KB	01_DataPrep_ExploratoryAnalysis
1_03_load_to_spark_s.ipynb	2/12/2018 10:01 AM	IPYNB File	5 KB	01_DataPrep_ExploratoryAnalysis
1_04_exploratoryDataAnalysis_Part1_s.ipynb	2/15/2018 8:21 PM	IPYNB File	33 KB	01_DataPrep_ExploratoryAnalysis
1_05_exploratoryDataAnalysis_Part2_s.ipynb	2/15/2018 8:21 PM	IPYNB File	38 KB	01_DataPrep_ExploratoryAnalysis



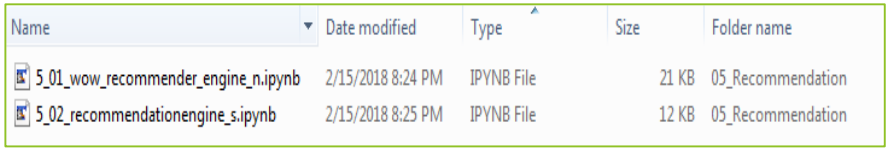
Name	Date modified	Type	Size	Folder name
2_01_spark_ml_wow_clusters_s.ipynb	2/15/2018 8:22 PM	IPYNB File	11 KB	02_Clustering



Name	Date modified	Type	Size	Folder name
3_01_TimeSeries_Forecasting_HoltWinters_s.ipynb	2/15/2018 8:22 PM	IPYNB File	6,783 KB	03_TimeSeries
3_02_Timeseries_Arima_s.ipynb	2/15/2018 8:22 PM	IPYNB File	1,148 KB	03_TimeSeries
3_03_Timeseries_Tensorflow_s.ipynb	2/15/2018 8:23 PM	IPYNB File	27 KB	03_TimeSeries

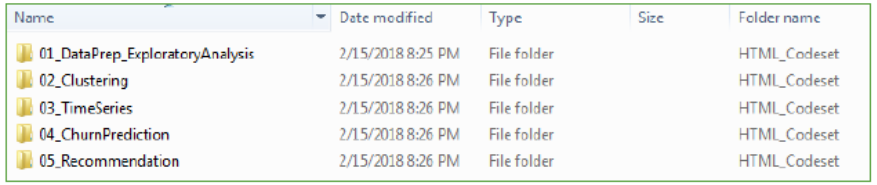


Name	Date modified	Type	Size	Folder name
4_01_wow_data_for_churn_create_d.ipynb	2/15/2018 8:23 PM	IPYNB File	30 KB	04_ChurnPrediction
4_02_analysis_wow_data_d.ipynb	2/15/2018 9:09 AM	IPYNB File	267 KB	04_ChurnPrediction
4_03_churn_featureengineering_s.ipynb	2/15/2018 8:24 PM	IPYNB File	36 KB	04_ChurnPrediction
4_04_churnprediction_s.ipynb	2/15/2018 8:24 PM	IPYNB File	259 KB	04_ChurnPrediction



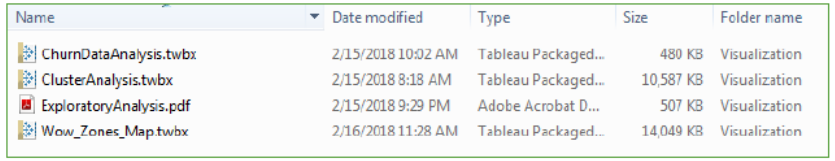
Name	Date modified	Type	Size	Folder name
5_01_wow_recommender_engine_n.ipynb	2/15/2018 8:24 PM	IPYNB File	21 KB	05_Recommendation
5_02_recommendationengine_s.ipynb	2/15/2018 8:25 PM	IPYNB File	12 KB	05_Recommendation

HTML Files



Name	Date modified	Type	Size	Folder name
01_DataPrep_ExploratoryAnalysis	2/15/2018 8:25 PM	File folder		HTML_Codeset
02_Clustering	2/15/2018 8:26 PM	File folder		HTML_Codeset
03_TimeSeries	2/15/2018 8:26 PM	File folder		HTML_Codeset
04_ChurnPrediction	2/15/2018 8:26 PM	File folder		HTML_Codeset
05_Recommendation	2/15/2018 8:26 PM	File folder		HTML_Codeset

Tableau Visualizations



Name	Date modified	Type	Size	Folder name
ChurnDataAnalysis.twbx	2/15/2018 10:02 AM	Tableau Packaged...	480 KB	Visualization
ClusterAnalysis.twbx	2/15/2018 8:18 AM	Tableau Packaged...	10,587 KB	Visualization
ExploratoryAnalysis.pdf	2/15/2018 9:29 PM	Adobe Acrobat D...	507 KB	Visualization
Wow_Zones_Map.twbx	2/16/2018 11:28 AM	Tableau Packaged...	14,049 KB	Visualization

Project in Github with all of the above code:

GitHub Repository

Thank you ☺