Data Scientist Test

It’s a standard part of our recruitment process to issue a test to job applicants. This three question test is designed to help us understand your quantitative analysis skillset, your perspective on industry objectives and your approach to solving challenges specifically related to game titles.

Your methodology is as important as what your answer turns out to be.

# 

# Question One

The following dataset (Dataset1) has been collated by your studio business team to allow the studio to better understand title purchase and gameplay trends. You have been asked to take this dataset and compile a one-page report for the entire studio’s benefit.

**1a. At a summary level, please describe the form and content of this report.**

Dataset1 is a CSV file that summarize the statistics at daily level. The report is compiled for three months starting from October 1st to December 30th.

Following table shows the description of each column

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Sample Data** | **Description** |
| Date | 1/10/14 | Date |
| Daily Installs | 153 | Total daily installs for the specified date |
| Weekly Installs | 153 | Installs for last seven days (The naming is poor - it is not weekly; it is for last seven days);  There are inconsistencies in the calucaltions here. Sometimes the totals are for last eight days and some times it is for 7days.  Initially the totals were computed for eight days and it was fixed after 25/10/14 to correctly count for last 7 days |
| Mean Install Duration | 47.71 | Average amount of time it took to install the application |
| Daily Unique Users | 921 | Daily unique users (Note could be more than installs because more users could use one install) |
| New Unique Users | 61 | New users (always less than or equal Daily unique users) |
| Sessions | 1284 | Total number of sessions created for the day (> unique users) |
| Sessions per User | 1.39 | Avg. sessions per user |
| Weekly Active Users | 921 | Weekly active users (cumulative total for the week) |
| 7D New Unique User | 61 | Weekly new users count |
| Mean Session Length | 127.01 | Avg. length of the session |
| Total Daily Play Time (Minutes) | 163092.27 | Total play time |
| LTD Play Time (Minutes) | 662928324 |  |
| 1 Day Returning Users | 674 | Returning users count After one day |
| 3 Day Returning Users | 56 | Returning users count After three days |
| 7 Day Returning Users | 283 | Returning users count After 7 days |
| 30 Day Returning Users | 543 | Returning users count After 30 days |
| Rounds Played | 467 | total rounds played |
| Players per Round | 4.64 | Total players per round |
| Hero Win Rate | 0.52 | Winning rate by leads |
| Creatures Killed | 21701 | clear and it is as described |
| Gold Spent | 112478 | clear and it is as described |
| Gold Earned | 52192 | clear and it is as described |
| Chests Opened | 1486 | clear and it is as described |

The descriptive stats for each feature along with interactions are

**1b. Please describe how you would go about creating this report. Valuable detail would include tools used, techniques used, intended functionality and how the information would be conveyed.**

Since this is created for Microsoft, clearly the data collected are using Asimov telemetry data captured in Cosmos.

At a higher level of architecture:

* Create a temporary hourly stream that extracts telemetry data for the games into hourly streams and its associated workflow in Cosmos. The location of the stream is   
  BASE + YYYY/MM/DD/HH   
  where BASE is the base location for all the streams for this project  
  YYYY/MM/DD/HH are GMT date formatted as such to stream the hourly data.
* Each new hourly stream is summarized into continuous daily stream;
* End of the day (GMT) daily streams are created for all the games.

From the above stream, for each game a daily summary is created by grouping last seven days of data.

I would SQLIZE the summary stream for Power BI reports that can be generated and fed to dashboards by analysts.

**1c. Assuming a long-term commitment to a report of the form you specified above, are there any changes would you make over time?**

A few suggestions I could think of are:

* Translate the numbers into charts and dash boards (Not exceeding more than concerned 4 KPIs)  
  For example, one KPI could be: Line chart on Daily Installs, Weekly installs, Time engaged, Weekly active users; Provide segmentation by other metrics.
* I suspect there are highly correlated variables such as, “Daily Unique Users” to “Total Daily Play time”; find such correlated features and translate them to identify the minimum set of KPI’s
* Monitor the above said correlations for any diversions and identify other features that are missing or revisit identification of KPIs
* Compare the numbers to three other gameplay that are selected from bottom 20% average and top 80%. This study can show how this gameplay compares to other gameplays
* Re-compute the weekly active users to remove inconsistency in the counting
* Add users leaving the (in addition to returning users) to compute churn
* Add users profile (age, gender, location, time, week day, seasons (summer, winter), current weather condition at the location, device, etc.– I omit at least 20 more dimensions I could think.) to segment the play time into more dimensions

# Question Two

Game teams at your studio have identified that monitoring how long players continue to play your most recent title is an area of particular interest. The studio is looking to you for guidance on how to manage information around this critical issue.

**2a. What recommendations would you give to the Business Manager regarding player engagement monitoring?**

**2b. What information does Dataset 1 contain which may be useful in regards to tracking long-term player engagement?**

I assume Long term player engagement column *(LTD Play time in Minutes)* contains the relevant data.

**2c. Using Dataset1, create an initial assessment of long-term player engagement challenges, issues and key data points. Please also provide a brief description of the techniques used.**

**2d. At a high level, assess the utility of your initial assessment.**

# Question Three

Your studio’s leadership team has requested you utilise a set of “Superdata” (Dataset2) to generate revenue predictions for an early title concept; the title in question being an FPS game with CCG and MOBA elements, planned for release on console, tablet and PC. Even an approximate prediction will enable the leadership team to make an informed decision regarding subsequent investment into the concept.

**3a. Generate an initial 18-month revenue prediction for this title. Please provide a description of your methodology and any relevant materials.**

**3b. Would you make any comments or observations regarding your revenue prediction?**