|  |
| --- |
| Summary Driver Analytics |
| Data Cleaning |

Revision Control

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Version No | Date | Change / Activity | By | Approver | Circulation |
| PA1 | 24-07-2019 | Initial Draft | Saurabh | Rajeev | Internal |
| A |  |  |  |  |  |

Reference Documents

|  |  |  |  |
| --- | --- | --- | --- |
| S No | Document Name | Version / Date | Owner |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Copyright: SENSORISE 2018

SENSORISE Confidential and Proprietary Information

All Rights Reserved

The contents of this document are restricted for circulation; please use pursuant to company instructions and the terms of the Non-Disclosure Agreement.

Table of Contents

**1** **Introduction 3  
2Table Of Contents 4**

# Introduction

* This document covers all snapshot of activities done so far right from data extraction to data cleaning for data to be processed in particular format for analysis
* Since there are numerous steps which are followed for the extracted process and cleaning steps are checked, edited and mapped properly in a tabular format
* For convenience following steps can guide to extract and perform analysis for data extraction and cleaning
* For ease of use and independence of environment folders needs to be created by the user and path needs to be specified where data is to be extracted or cleaned
* For generic purpose Directory (**dir**) is specified in all main files and python scripts where a user can specify the directory and make named folders to get the data stored
* Each and every program has been optimized with highly efficient coding mechanism
* The code needs to be run in cmd prompt as each and every code is multiprocessing four or more files together

**Note: Python IDE does not support multiprocessing**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | Program | Input | Output | Purpose | Description |
|  | IMEI Extraction | Data from Port Files  in specified directory(dir) of  “Clean Records//  Devicedata//” folder | Extracted Data in directory of Clean Records containing four main subfolders:   * Data –: IMEI Files * Error –: Bad Records * FileSummary -: Summary records * Track -: Good Records   60 days ETA | The purpose of this program is to extract IMEI files having proper validated fields so that vehicle wise analysis could be done  i.e Each vehicle has an unique IMEI | 1. Create **Clean Records** folder in user specified directory (dir) 2. Make five folders in the subdirectory \\**Clean Records** :    1. Devicedata    2. Data    3. Error    4. FileSummary    5. Track 3. In Python script folder IMEI Extraction\\ run the main file specifying directory in it( dir = “”) 4. Record\_Parse\_Modified.py file is linked with the main file |
|  | IMEI Cleaning | Data after IMEI Extraction  in specified directory(dir) of  “Clean Records// Data//” folder  (1 day ETA) | * Extracted Data in directory of Clean Records\\IMEI\_Files containing all IMEI files which are cleaned and extracted * “Clean\_Record\_Summary.csv”   File comprising records at each IMEI level  \*\*Number of files may not be same and less as IMEI files having all duplicate records and one record will be omitted\*\* | * Remove all duplicate , redundant records * Mark trips based on time interal * Find spurious points and clean up with the algorithm to find segments in trips * Remove all those points in paths which have minimum tolerance in distance and speed | 1. Make IMEI\_Files in the subdirectory \\**Clean Records** : 2. In Python script folder IMEI Cleaning\\ run the main file specifying directory in it( dir = “”) 3. Three python files are linked with this main file: 4. Mark\_Point 5. Separate\_Segments 6. RDP\_Modified |
|  | IMEI Heuristics | Data after IMEI Cleaning  in specified directory(dir) of  “Clean Records//IMEI\_File//” folder  (1.5 day ETA) | Extracted Heuristics Data in directory of Clean Data\\IMEI\_Heuristics which contain 3 files:   * Distance Heuristics: distance and frequency * Speed Heuristics: Speed and frequency * Time Heuristics: Time and frequency | The role of this program is to get insight about distance, speed and time distribution of the data i.e one of the useful criterion about vehicle movement | 1. Make **IMEI\_Heuristics** folders in the subdirectory \\**Clean Records** 2. In Python script folder IMEI Cleaning\\ run the IMEI\_Heuristics.py specifying directory in it( dir = “”) |
|  | Summary Screen and  Summary Visualization | Data after IMEI Cleaning  in specified directory(dir) of  “Clean Records//IMEI\_File//” folder  (1 day ETA) | * Extracted Description Data for visualization in directory of Clean Records\\IMEI\_Description which contain 7 files: * Days\_Description * Distance\_Description * Hour\_Description * IMEI\_Description * Segment\_Description * Time\_Description * Trip\_Description   \*\*Please copy paste Clean\_Record\_Summary.csv  File in the IMEI\_Description folder for visualization\*\*   * Summary visualized screen depicting visually of the summary of the IMEI description | Summary screen displays information about descriptive study i.e where insights could be generated about the pattern of day, day of week , hour of day, total trips in IMEI’s etc. and other important criterion about PVT points | 1. Make **IMEI\_Decription** folders in the subdirectory \\**Clean Records** 2. In Python script folder IMEI Cleaning\\ run the IMEI\_Description.py specifying directory in it( dir = “”) 3. After description files are generated Please copy paste Clean\_Record\_Summary.csv   File in the dir \\ **Clean Records** \\**IMEI\_Description** folder for visualization   1. In Python script folder IMEI Cleaning\\ run the Summary\_Visualization.py   specifying directory in it( dir = “”) |
| 5. | Feature Extraction | Data after IMEI Cleaning  in specified directory(dir) of  “Clean Records//IMEI\_File//” folder  (3 Days ETA) | * Extracted Data Features for model building in directory of Clean Records\\Features which contain 3 files: * Feature\_IMEI * Feature\_Segment * Feature\_Stretch | Feature extraction is an extensive program where we can have generated all important criterions for model building in each Segment, Trip and IMEI level | 1. Make **Featured** folders in the subdirectory \\**Clean Records** 2. In Python script folder Feature Extraction\\ run the Main\_File\_Features.py specifying directory in it( dir = “”) 3. Data\_Features is linked with the file for extracting features at each level |
|  | Feature Extraction | Data after IMEI Cleaning  in specified directory(dir) of  “Clean Records//IMEI\_File//” folder  (3 Days ETA) | The following files are generated after processing with IMEI files and validating start and stop points   * Summary\_Cleaned\_IMEI * Geo\_IMEI * Centroid * Start\_Cluster * End\_Cluster * Cluster\_Features | Here we generated files with proper start and stop points ( validated ) to get path of travel of a vehicle  The paths are basically routes which we need to process for our driver behaviour | 1. In features foldres six .csv files will be generated mentioned in output 2. Just Run the script “Segment\_Stretch \_Features” by mentioning the **directory (dir)** |