

## **DATA MANAGEMENT**

# **Sensor Data** **Used Car Database**

Report By

Yashawant Prabhakar Parab

## Task B

To perform Task B, 2 datasets were used.

- 1) **Sensor Data**
- 2) **Used Car Database**

Tools used for this task are:

- 1) Google Cloud Platform Trifacta
- 2) Python
- 3) Raspberry Pi Model B (HC-SR04 Distance Sensor, RaspbianOS and all Dependent tools)
- 4) Microsoft Excel

### Sensor Data

The sensor data is basically the distance of other objects from the object on which the sensor is mounted to avoid collision. The raw data is extracted from the Distance sensor module which measures the distance and stores it in the log file. To execute this, Raspberry pi, HC-SR04 Distance Sensor and some dependent tools were used.

#### **Dimension:**

The collected data was in raw format and met the data quality dimension. Some operations were performed to unclean this data. Data is in the form of positive numeric value which. Also, date and time is added in this dataset to make it messier.

Sr. No	Data Element
1	CreatedDate
2	CreatedTime
3	Distance in Cm

#### **1) Data Element: CreatedTime and CreatedTime**

Dimension: Completeness, Validity, Timeliness

##### **Steps:**

- a) To make data messy Microsoft excel tool was used. The date format was converted into number by using format cell function. Also, the same operation was used on CreatedTime column.
- b) By performing this operation, the accuracy of the dataset was ruined

#### **2) Data Element: Distance in Cm.**

Dimension: Completeness

- a) The captured data was in positive numeric value. To make data, inaccurate some rows was converted to negative numeric value.

- b) Also, as per the specification of distance sensor, the highest measurement of the sensor is 400 cm but in the considered dataset, there were few values which were exceeding this limit.

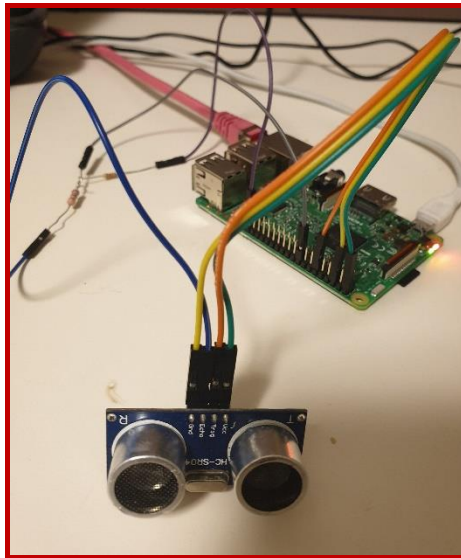
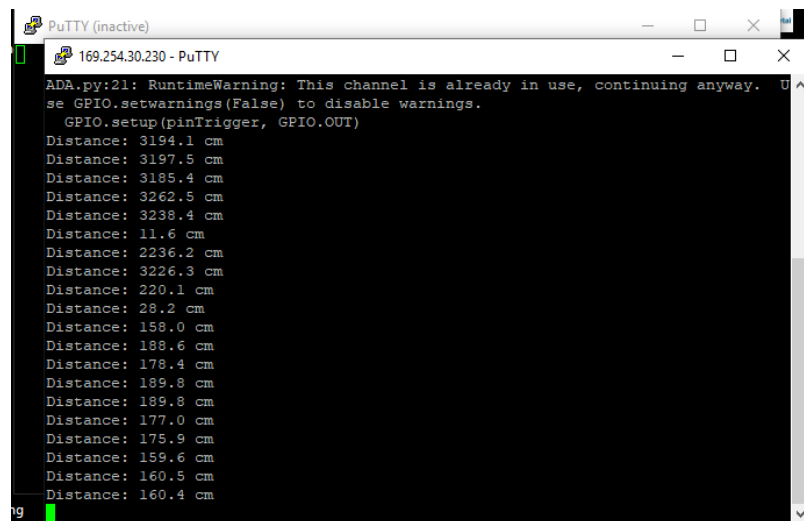


Fig: 1



```
169.254.30.230 - PuTTY
ADA.py:21: RuntimeWarning: This channel is already in use, continuing anyway. Use
GPIO.setwarnings(False) to disable warnings.
  GPIO.setup(pinTrigger, GPIO.OUT)
Distance: 3194.1 cm
Distance: 3197.5 cm
Distance: 3185.4 cm
Distance: 3262.5 cm
Distance: 3238.4 cm
Distance: 11.6 cm
Distance: 2236.2 cm
Distance: 3226.3 cm
Distance: 220.1 cm
Distance: 28.2 cm
Distance: 158.0 cm
Distance: 188.6 cm
Distance: 178.4 cm
Distance: 189.8 cm
Distance: 189.8 cm
Distance: 177.0 cm
Distance: 175.9 cm
Distance: 159.6 cm
Distance: 160.5 cm
Distance: 160.4 cm
```

Fig: 2

## Further Analysis

To make this Raspberry pi model more flexible Google Cloud Platform was used. In this GCP, there is a dedicated option to connect IOT device on cloud.

## Used car sensor

### Data elements

1	Crawled_Date
2	Crawled_Time
3	Price
4	Vehicle Type
5	PowerPS
6	Model
7	Kilometer
8	FuelType
9	Data and Time Creation
10	Last seen

To unclean the data, several operations were performed on the data elements mentioned above.

#### 1) Data Element: Crawled Date, Crawled Time, last seen, Data and Time Creation

- a) Date format operation was performed on these data elements using the formula given below. The separators were removed from the columns.

- `DATEFORMAT (dateCrawled, 'EEE,MMddyyyy-HH:mm:ss')`
- `/ [100]: [0-9] [0-9] :/`

- b) The accuracy, validity, Completeness of the dataset were ruined.

#### 2) Data element: Price

- a) Values which were less than 250 were converted to 0 to make the data inaccurate. The price of the car cannot be 0 or negative.

#### 3) Data element: Vehicle Type

- a) The blank value of the vehicle type was replaced to dugout boat which is not a type of vehicle. Using this operation, the conformity of the data element was broken.

#### 4) Data element: PowerPS

- a) Car power value was misplaced with the kilometre data column, those power values were less than 20. With this process timeliness data quality dimension was converted into incorrect dimension

#### 5) Data element: Model

- a) In Model, there were few missing values which were corresponding to the Car model. Automatic was added as a model in null value.
- b) Performing this operation, some inaccuracy, inconsistency of data was introduced. This makes the data invalid and failed at conformity check.

#### 6) Data element: Kilometre

- a) Values less than 90000 were replaced with negative numeric value. The column states car kilometer in positive numerical value. But in this case car had negative kilometre value which is inaccurate and invalid

#### 7) Data element: FuelType

- a) Data contain all the fuel type present in the automobile industry. Some data had null values that were replaced to Hydrogen which is not a fuel. This makes the data inconsistent, inaccurate, invalid