Experiential Learning
Connected Vehicles Applications

Created by _______Lin Lyu & Lifei Zhu________

Table of Contents

Background of Problem

The U.S. Department of Transportation's (USDOT's) Connected Vehicle program is working with state and local transportation agencies, vehicle and device makers, and the public to test and evaluate technology that will enable cars, buses, trucks, trains, roads and other infrastructure, and our smartphones and other devices to "talk" to one another. Connected vehicles could dramatically reduce the number of fatalities and serious injuries caused by accidents on our roads and highways. Connected vehicles have significant advantages over new technologies now appearing in high-end vehicles, such as radar, lidar, cameras, and other sensors. For one thing, connected vehicle technologies and applications have a greater range than on-board vehicle equipment, which will allow us to receive alerts of hazardous situations much earlier, providing more time to react and prevent an accident. Also, connected vehicle technology does not depend on "line of sight" communications to be effective, unlike radar. Connected vehicle technology is also less expensive to install than radar and camera equipment in vehicles. This will enable it to become standard equipment in the future on practically all vehicles, not just luxury cars.

At present, the social ownership of automobiles is maintained at a high level, and the demand for online monitoring of vehicle status information, remote positioning, and road navigation is gradually increasing; the increase in automobile functions and the continuous improvement of the level of intelligence have led to more and more vehicle structures[4]. Complex and detailed, with a wide variety of vehicle status information, the technical requirements for maintenance personnel are becoming more and more stringent, and higher requirements are also placed on online monitoring equipment. On the other hand, road safety evaluation is particularly important for improving road traffic safety. Overall, there were an estimated 263.6 million registered vehicles in the United States in 2015, most of which were passenger vehicles. 36,560 people were killed in traffic crashes in 2018, most of which were caused by driver's cognition and behavior decision errors. Therefore, the driver is also eager to know some data during the driving of his car, so as to manage the vehicle more actively. At present, the transportation field has become one of the key promotions and demonstration fields of the Internet of Things strategic technology industry, and the Internet of Vehicles is one of the future development trends of automobiles.

On board diagnostics (OBD) is an important part of vehicle networking and a modern vehicle emission inspection designed to detect computerized problems[1-2]. During an OBD I/M check, a technician will attach a cable to the vehicle's onboard computer to see if the vehicle's emissions equipment is damaged or otherwise malfunctioning. The check can assist in preventing major engine problems and therefore major expenses caused by failing equipment while subsequently benefiting the environment in reducing the amount of harmful gasses emitted.

And nowadays different states have different test frequencies, for example, in Colorado, an emissions test is required every two years for vehicles more than seven model years old while in Pennsylvania, a test should be done every year.

After we studied the emission test data of Pennsylvania, we found that many cars did not strictly comply with the requirement of conducting an emission test once a year. And different cars have different characteristics, such as odometer, model year, emission test interval, etc., which may affect the result of the successful detection of the car. So, we are trying to find out if there is any relationship between failure rate of OBD test and the frequency of testing while trying to explore how the various characteristics of the car affect the probability of the car passing the test.

The Internet of Things & Cloud-based Program and on-board diagnostics (OBD) of Connected Vehicles show a good potential for using OBD data to predict whether the vehicle is a high emitter or not. So, it is of great importance for us to learn and explore this relevant knowledge in these emerging research areas.

Literature Review

Brief History of PA's Emission Inspection Program

In 1990, Congress set the requirement for an enhanced program in the Clean Air Act Amendments. In October 1995, the Commonwealth announced Pennsylvania's decentralized Vehicle Emissions Inspection and Maintenance Program. In 1997, the Commonwealth initiated an enhanced auto emissions testing program that was designed and implemented to improve the air quality in nine counties in the Philadelphia and Pittsburgh regions. In November 1999, an expanded emissions program was to be implemented in the South Central and Lehigh Valley areas. In January 2000, the Stakeholders Groups made their recommendations. They suggested an emission program similar to the program already in place in the Pittsburgh region. In January 2001, faced with technology advances and other factors, PENNDOT and DEP announced the creation of a Vehicle Emissions Inspection Policy Review Group to determine the impact of the changes on state plans for an expanded vehicle emissions-inspection program. One of the key technology developments is the presence in new cars of on-board diagnostic technology. On-board diagnostics have the potential to be used for emissions tests that are both cheaper and easier. In May 2003, PENNDOT and DEP announced an agreement to settle the pending lawsuits over Pennsylvania's program for automobile emissions testing. The changes called for in this settlement agreement were to bring the state's emissions testing program into compliance with federal air quality standards and remove the threat of federal sanctions, while having a minimal effect on most of the state's drivers. As of June 30, 2004, all 25 counties will have fully implemented a vehicle emission testing program.

The state of Pennsylvania requires vehicle emissions tests and safety inspections for non-exempt vehicles in getting their initial vehicle registration or to complete an annual registration renewal in eligible counties. The Pennsylvania Department of Transportation (PennDot) sends out testing notices to owners indicating when the test must be completed by. Emissions tests can be completed up to 60 days before the testing deadline.

OBD in connected vehicles

Overview of OBD

OBD stands for On-Board Diagnostic. It is the standardized system that allows external electronics to interface with a car's computer system. It has become more important as cars have become increasingly computerized, and software has become the key to fixing many problems and unlocking performance.

OBD systems installed on vehicles included, among other things, an engine control module that monitored the engine controls and emission related components, a malfunction indicator lamp (MIL) located on an instrument panel and other supporting circuitry and memory[3]. When a malfunction was detected by the OBD system, the MIL illuminated to provide notice to the vehicle operator of an engine or emissions malfunction. At the same time, the OBD system stored in memory the DTCs corresponding to the specific malfunction detected.

Remote OBD

There is a trend about implementing a Remote On-board Diagnostic (OBD)

Inspection/Maintenance (I/M) program either as an addition to an existing periodic inspection system or as a stand-alone program[5].

Remote OBD offers the opportunity for owners of OBD-equipped vehicles to avoid having to get a periodic, physical inspection of their motor vehicle, and yet still comply with the requirements for an OBD inspection. Due to the continuous nature of Remote OBD, as opposed to the annual or biennial tests now done in periodic I/M programs, Remote OBD can find problems with motor vehicles more quickly, leading to faster repairs and, thus, more emission benefits.

A Remote OBD system consists of three basic elements: The first one is a Remote OBD Link, which is a device on-board the vehicle that continuously captures data from the OBD system and transmits it wirelessly. The second one is a method of receiving the data from the Remote OBD Link, which may be an existing wireless network, or one specifically created for the purpose. The third one is a Data Management System (DMS) that processes the Remote OBD data and performs I/M specific functions including determining pass/fail conditions and reporting to the administrative agencies and motorists. The Workgroup has attempted to address many of the technical and policy questions related to implementing Remote OBD. One goal was to make it easier for states to add a Remote OBD program to Requests for Proposals. Another goal was to standardize to the extent reasonable the approach to Remote OBD such that systems and technologies may be compatible across state lines.

In broad terms, a Remote OBD program will involve either installing a device on a 1996 or newer OBD equipped vehicle or using existing on-vehicle telematics to monitor the OBD system on a continuous basis. When a problem occurs with the vehicle’s emission control system and the MIL is “commanded on, “the motorist will be notified that vehicle repair work is required. In such cases, the program will take steps to follow-up on such vehicles to ensure that the MIL goes “off” within the grace period allowed for repairs. This approach frees the motorist from having to show up for a periodic inspection of the vehicle, and might be substantially less expensive than the traditional periodic inspection program.

Two questions need to be considered in remote OBD

One of the basic questions regarding Remote OBD is what constitutes continuous monitoring. The following working definition is recommended and used as the basis for this guidance: A Remote OBD system should be designed such that 80% of the vehicles enrolled in the program and operated routinely in the area communicate with the network at least once every 14 days, with the other 20% “seen,” i.e., they communicate with the network, at least once per month. It is expected that vehicles will drop out of the program as they are transferred out of the area or scrapped, and that such withdrawals should be excluded from this calculation. If after 2 months a vehicle is not seen by the Remote OBD system, the motorist should be contacted to determine the cause. Action should be taken as appropriate to either remove non-respondents, transferred vehicles, and scrapped vehicles from the DMS or to fix any problem that might prevent the required transmission of data from the vehicle to the network. A vehicle that is routinely “seen” according to this standard would be considered in compliance with the information reporting standards and eligible for re-registration.

Another basic question relates to the conditions under which a motorist participates in

Remote OBD. The Workgroup believes that states may, at least initially, take a voluntary approach to implementing Remote OBD, allowing motorists to opt into the program and out of periodic inspection. The Workgroup recommends that motorists that wish to opt in to Remote OBD enter into an agreement with the I/M jurisdiction. In such an agreement, the motorist will:

•Agree To abide by the terms and conditions of the Remote OBD program.

•Agree not to tamper with or destroy the Remote OBD device.

•Commit to timely servicing of the vehicle after being notified by the I/M program that repairs are needed. If a motorist fails to comply with the repair requirements, then the program will need to take enforcement action to ensure that compliance is achieved. This may be accomplished by removing the vehicle from the Remote OBD program and immediately scheduling it for a periodic inspection. Alternatively, license or registration suspension can be used to bring motorists into compliance.

•Commit to be "seen" in the I/M area and, if necessary, intentionally drive by a monitor periodically so that the OBD system can transmit required data.

Internet of Things

IoT is a network constituted by uniquely identifiable commodity objects or devices equipped with some sensing system[6]. IoT paradigm promotes a seamless amalgamation between the smart devices, scatter around us, and the physical world to ensure full automation that eventually ameliorates human life. Some of the examples of IoT-enabled commodity devices or things include heart monitoring implants, automobiles with embedded sensors, firefighter devices, smart thermostat systems, and Wi-Fi enabled washer/dryers. As the arena of IoT is expanding, the number of IoT-enabled applications is also rapidly growing, which results in massive growth of smart devices in multiple orders comparatively[10]. This swift increase in the number of sensing things is responsible for generating and storing a plethora amount of Big Data at a much faster rate. The data needs to be engineered and analyzed, which require robust services and analytical systems as the traditional techniques are unable to successfully and efficiently accost such data stress[11].

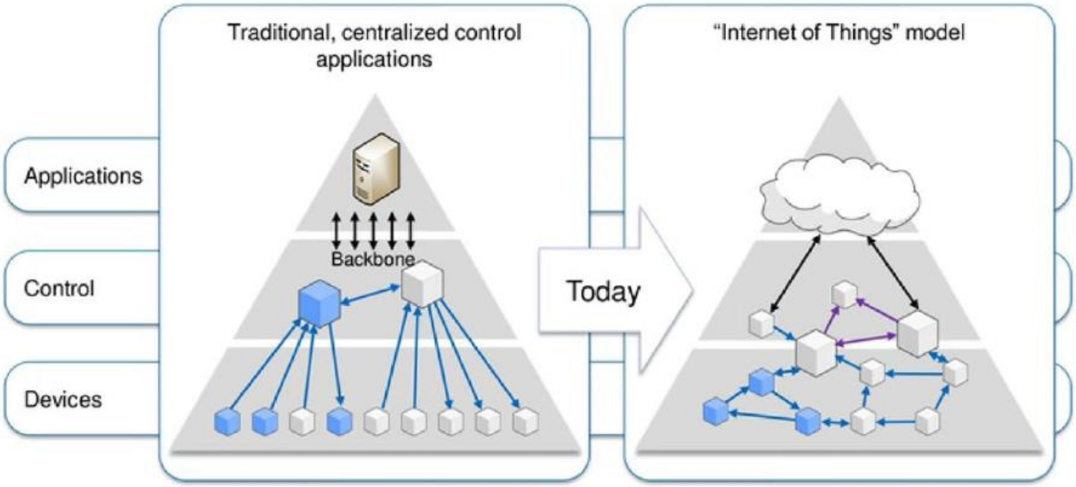

Traditional, centralized control applications
"Internet of Things" model
Applications
Control
Devices
Backbone
Today

Figure 1 Traditional versus IoT paradigm
(Source: http://www.db.in.tum.de/teaching/ws1314/industrialIoT/)

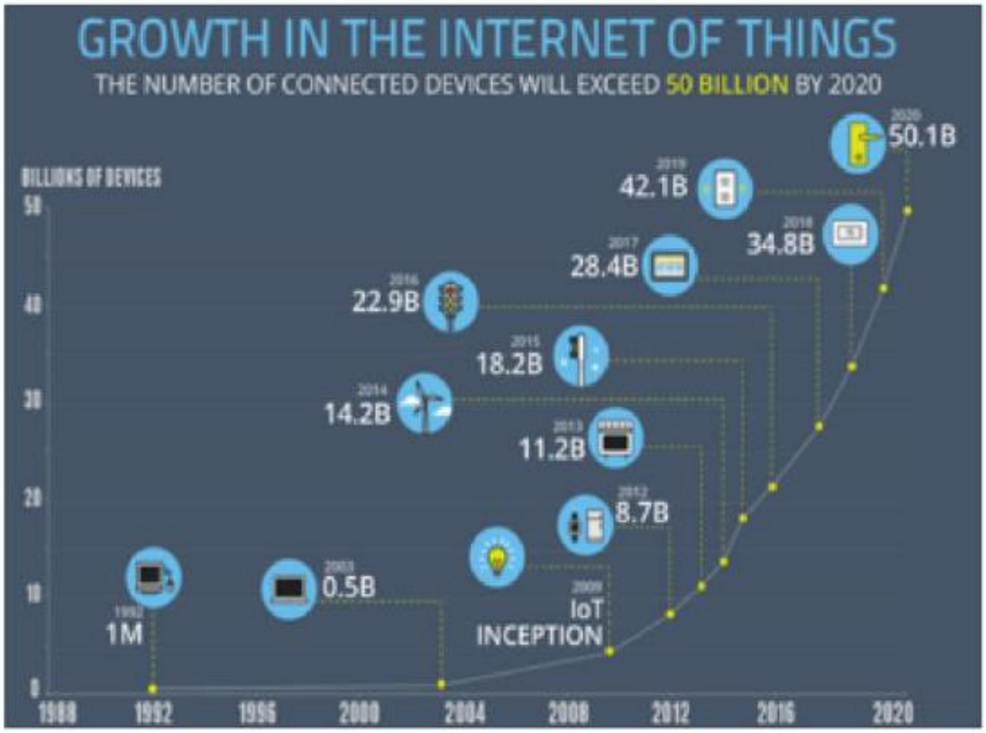


Figure 2 IoT growth over years
(Source: www.ncta.com/broadband-by-the-numbers/)

The emerging services and analytics advocate the service delivery in a polymorphic view that successfully serves a variety of audience. The amalgamation of numerous modern technologies such as cloud computing, Internet of Things (IoT) and Big Data is the potential support behind the emerging services Systems. Today, IoT is taking the center stage over the traditional paradigm. The evolution of IoT necessitates the expansion of horizon to deal with emerging challenges.

Cloud-based Program

Cloud-based vehicle analytics is a new endeavor that opens up a sludge of possibilities for providing superior functions and services in the transportation sector particularly for emissions testing, which may solve many, if not all, of the issues and concerns regarding emissions, performance, repairs, maintenance, driving behavior, accident reconstruction and others[7-10]. This requires the cooperation of all stakeholders including EPA, CARB, ASA, NHTSA, SEMA, OEMs and motorists[11]. The proposed solution involves real-time collection of all in-vehicle data and uploading it to the cloud in such a manner that all of what can be done at a dealer shop through OBD II, can be also done with the vehicle data available at the cloud[8]. The power, benefits and advantages of this solution must not be underestimated. Of course, this solution will only work if concerns regarding privacy, security, confidentiality, safety and individual rights are properly addressed[12-13].

Integrated cloud computing and IoT ecosystem

The growing smart communication among the things, especially the sensor equipped, under the purview of IoT is resulting into production of an incredibly large amount of Big Data[9]. Soon, the concept of IoT is going to deeply pervade through human life intending to automate the routine chores. The needs for the highly robust computing and analytical services to address and analyze the data-related challenges with the exceedingly reduced cost have further brought the service-rich cloud computing to forefront[14-15]. The rising interest in IoT and cloud provisioning systems has triggered the surge in cloud-housed comprehensive analytics.

Academia, industry and research communities are rapidly developing robust analytic frameworks and housing them in cloud environments, and are delivered and utilized as analytics-as-a-Service. In 2008, to deal with the collaborative complex computations, instead of data in the cloud, Cerri et al. devised an analytical framework in the cloud, called knowledge in the cloud that delivered multifarious knowledge[16]. Figure 3 shows the cloud computing and IoT ecosystem. It can be noticed that a sensing system, may differ in attributes, is embedded into the objects/devices around us. The device communication with the cloud is facilitated through the robust cloud services. However, in IoT infrastructure a smart connectivity gateway, may be with the existing networks, is required to further strengthen the device communication. Therefore, the communication between the cloud and the external word is feasible only through robust and reliable cloud services[17-20].

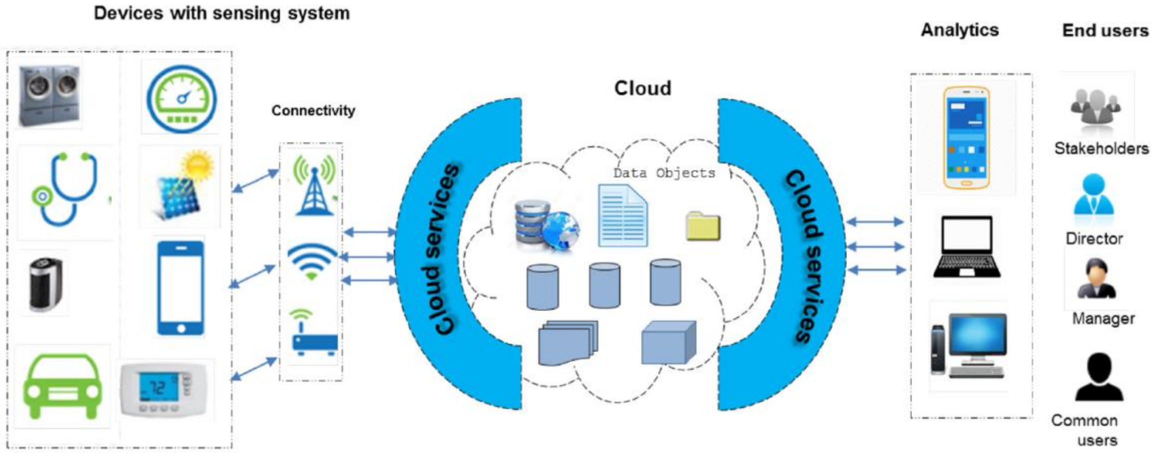


Figure 3 Cloud computing and IoT ecosystem

Research Questions Motivated from Literature Review and Database

Generally, our task focused on exploring and studying the relationship between OBD test frequency and test failure rate and using different models and methods to analyze. Our target is to see if we expand or shorten the OBD testing time interval, will there be any change of the OBD testing failure rate. Under these circumstances, we have tried three methods to analyze the relationship between test result and test time interval.

1. If we divide test frequencies of vehicles into different groups, we wonder if there is an obvious optimal time interval for OBD test.
2. Build an OLS model to see if there is any linear relation between test frequency and test result.
   $y=ax+b$
3. Logistic regression

Logistic regression is essentially linear regression. But Logistic regression is specifically used to classify 0/1 problems. Our emission test result is a 0/1 problem. So theoretically, using LR to fit our data to find that the relationship between the success of the emission test result and the time interval of the test will have a better effect.

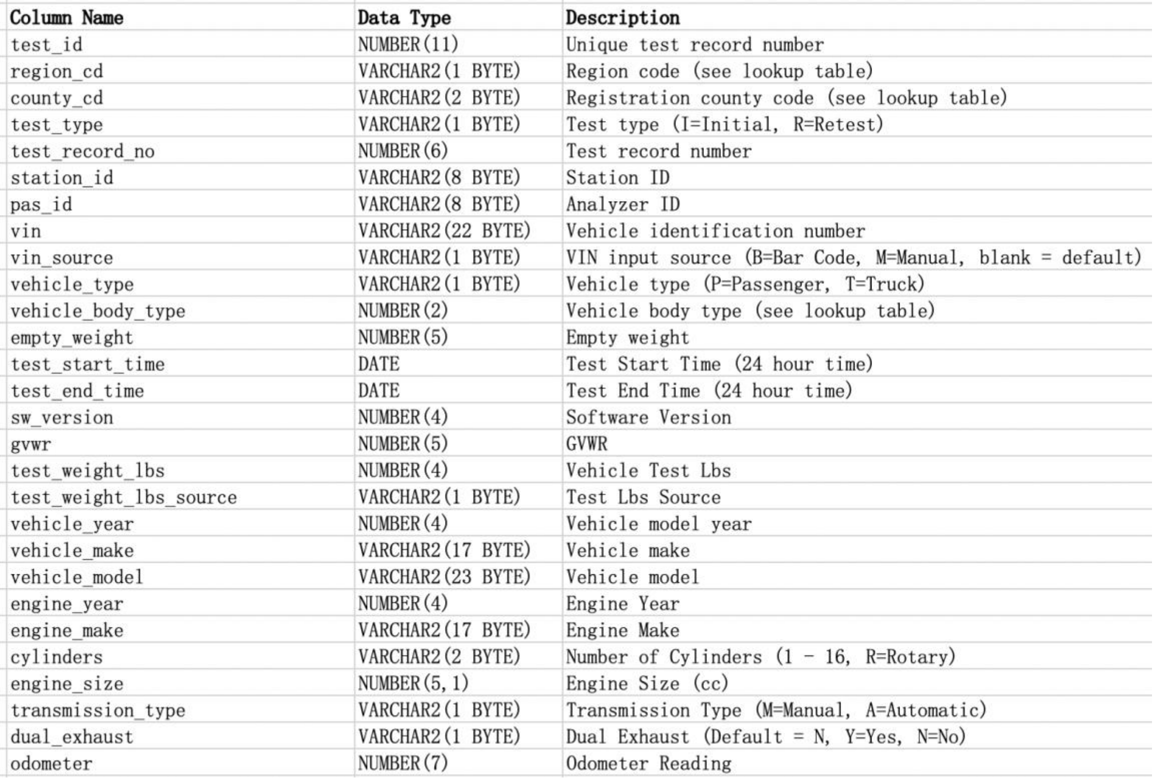
Research Process to Answer Questions

We choose emission data of Pennsylvania from 2000 to 2016. To obtain high quality emission data, we filter some “noise” and combine the data of all these years to create data dictionaries. Finally, we can do our data analysis and draw some conclusions.

Data Used

Table 1 is just a part of the whole data dictionary of our emission data.

Table 1 Data Dictionary of Emission test in Pennsylvania



Column Name	Data Type	Description
test_id	NUMBER(11)	Unique test record number
region_cd	VARCHAR2(1 BYTE)	Region code (see lookup table)
county_cd	VARCHAR2(2 BYTE)	Registration county code (see lookup table)
test_type	VARCHAR2(1 BYTE)	Test type (I=Initial, R=Retest)
test_record_no	NUMBER(6)	Test record number
station_id	VARCHAR2(8 BYTE)	Station ID
pas_id	VARCHAR2(8 BYTE)	Analyzer ID
vin	VARCHAR2(22 BYTE)	Vehicle identification number
vin_source	VARCHAR2(1 BYTE)	VIN input source (B=Bar Code, M=Manual, blank = default)
vehicle_type	VARCHAR2(1 BYTE)	Vehicle type (P=Passenger, T=Truck)
vehicle_body_type	NUMBER(2)	Vehicle body type (see lookup table)
empty_weight	NUMBER(5)	Empty weight
test_start_time	DATE	Test Start Time (24 hour time)
test_end_time	DATE	Test End Time (24 hour time)
sw_version	NUMBER(4)	Software Version
gvwr	NUMBER(5)	GVWR
test_weight_lbs	NUMBER(4)	Vehicle Test Lbs
test_weight_lbs_source	VARCHAR2(1 BYTE)	Test Lbs Source
vehicle_year	NUMBER(4)	Vehicle model year
vehicle_make	VARCHAR2(17 BYTE)	Vehicle make
vehicle_model	VARCHAR2(23 BYTE)	Vehicle model
engine_year	NUMBER(4)	Engine Year
engine_make	VARCHAR2(17 BYTE)	Engine Make
cylinders	VARCHAR2(2 BYTE)	Number of Cylinders (1 - 16, R=Rotary)
engine_size	NUMBER(5,1)	Engine Size (cc)
transmission_type	VARCHAR2(1 BYTE)	Transmission Type (M=Manual, A=Automatic)
dual_exhaust	VARCHAR2(1 BYTE)	Dual Exhaust (Default = N, Y=Yes, N=No)
odometer	NUMBER(7)	Odometer Reading

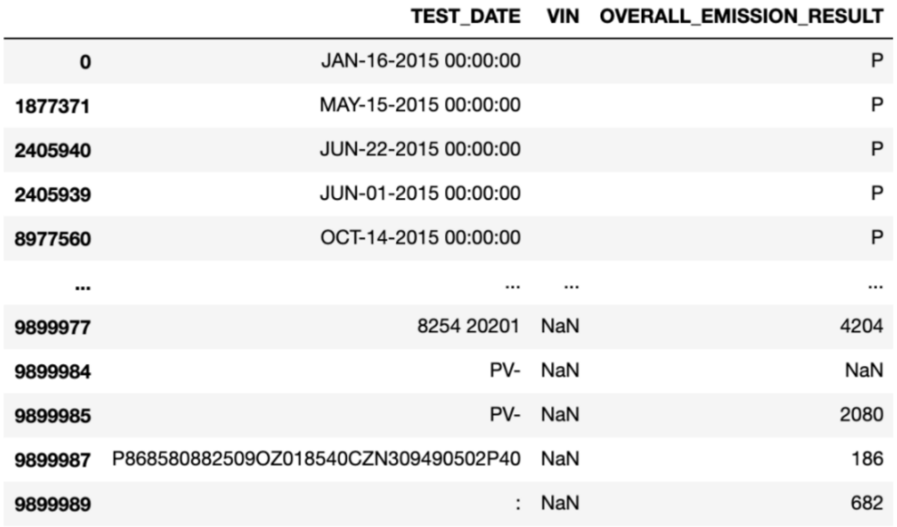
We are using emission inspection data from year 2000 to year 2016 in Pennsylvania and one month has one csv file. We only read in four columns: ‘VIN’, ‘TEST_DATE’, ‘OVERALL_EMISSION_RESULT’, ‘EMISSION_TEST_TYPE_ID’, which are the most important columns for our goal, finding out how many ‘high quality’ emission data we could use and calculating the failurate.

For ‘OVERALL_EMISSION_RESULT’, there are three types: P = Pass, F = Fail, N= NA, and we count F and N are all fail records.

For ‘EMISSION_TEST_TYPE_ID’, there are three emission types: A=ASM, T = TSI, O=OBD, V=Visual. We just care about ‘EMISSION_TEST_TYPE_ID’ is ‘O’, which means cars have OBD tests. And some very previous years, like 2000 and 2001, almost all cars used ASM and TSI, like in 2000, there are 1691919 records of ASM and 1308884 records of TSI, no records of OBD.

Method/Approach and Results

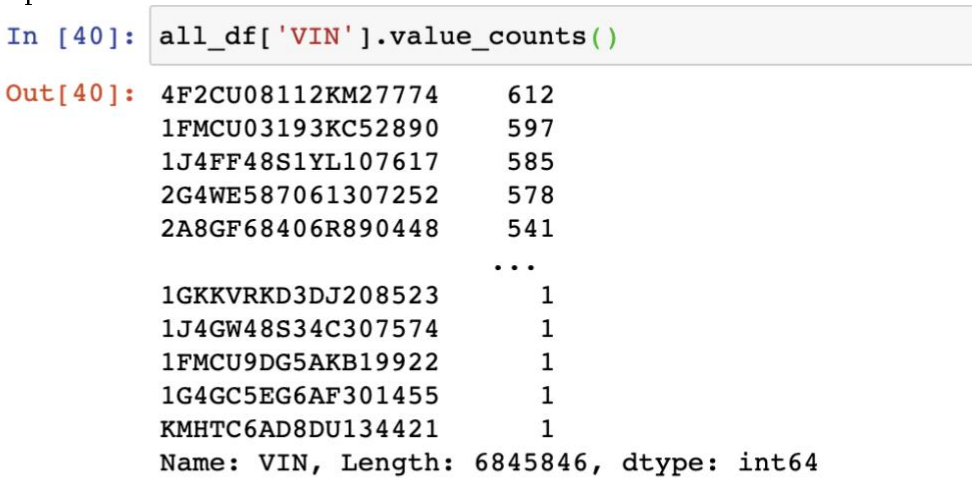
(1) Data introduction and data clean
At first, we just use year 2015 and year 2016 data, total 24 months, for having a quick look at the data, to see what information the data has, how to do data cleaning. Using the whole 17 years date to have a look would be time consuming, since the whole dataset maybe about 60GB.



	TEST_DATE	VIN	OVERALL_EMISSION_RESULT
0	JAN-16-2015 00:00:00		P
1877371	MAY-15-2015 00:00:00		P
2405940	JUN-22-2015 00:00:00		P
2405939	JUN-01-2015 00:00:00		P
8977560	OCT-14-2015 00:00:00		P
...	...	...	...
9899977	8254 20201	NaN	4204
9899984	PV-	NaN	NaN
9899985	PV-	NaN	2080
9899987	P868580882509OZ018540CZN309490502P40	NaN	186
9899989	:	NaN	682

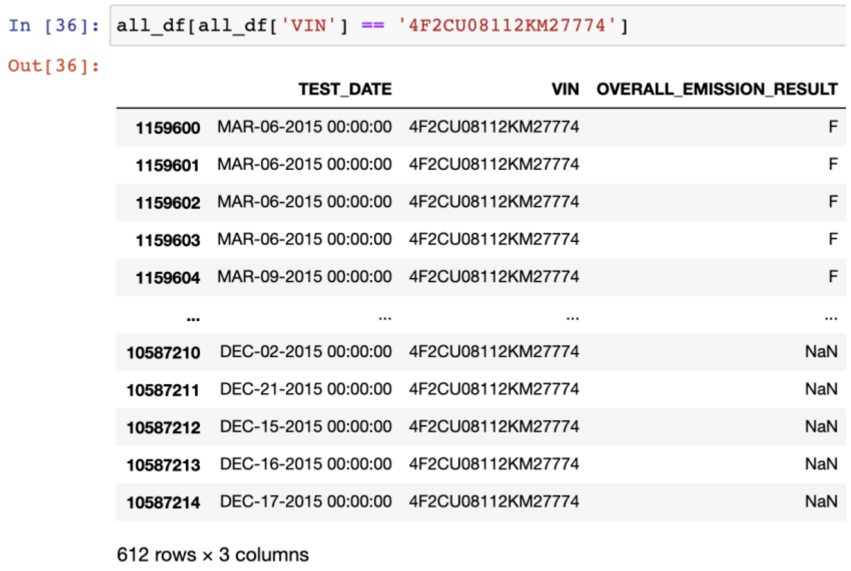
Figure 4 Raw Data of 2015 & 2016

After reading in the two year data, as we can see from the Figure 4 above, there are many ‘NaN’ values.
Then we found many cars do not do inspection on time. In Pennsylvania, cars are requested to do emission inspection once a year. But from the two year data, we found that if a car went to an inspection site in a specific month in 2015, we can not find an inspection record in this specific month in 2016. This situation may be because some car owners may forget to do inspection on time or they have something else to do resulting in a delay of inspection.



```
In [40]: all_df['VIN'].value_counts()

Out[40]: 4F2CU08112KM27774    612
         1FMCU03193KC52890    597
         1J4FF48S1YL107617    585
         2G4WE587061307252    578
         2A8GF68406R890448    541
                             ...
         1GKKVRKD3DJ208523      1
         1J4GW48S34C307574      1
         1FMCU9DG5AKB19922      1
         1G4GC5EG6AF301455      1
         KMHTC6AD8DU134421      1
         Name: VIN, Length: 6845846, dtype: int64
```

Figure 5 Times of Test in 2015 & 2016

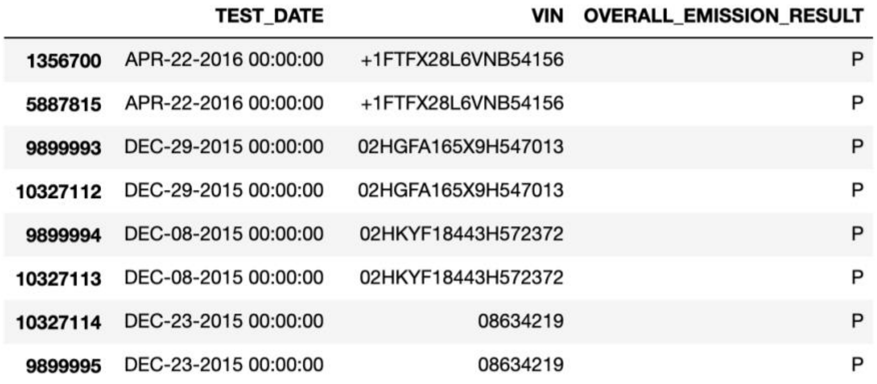


```
In [36]: all_df[all_df['VIN'] == '4F2CU08112KM27774']

Out[36]:
                    TEST_DATE                VIN  OVERALL_EMISSION_RESULT
 1159600  MAR-06-2015 00:00:00  4F2CU08112KM27774                        F
 1159601  MAR-06-2015 00:00:00  4F2CU08112KM27774                        F
 1159602  MAR-06-2015 00:00:00  4F2CU08112KM27774                        F
 1159603  MAR-06-2015 00:00:00  4F2CU08112KM27774                        F
 1159604  MAR-09-2015 00:00:00  4F2CU08112KM27774                        F
     ...                   ...                ...                      ...
10587210  DEC-02-2015 00:00:00  4F2CU08112KM27774                      NaN
10587211  DEC-21-2015 00:00:00  4F2CU08112KM27774                      NaN
10587212  DEC-15-2015 00:00:00  4F2CU08112KM27774                      NaN
10587213  DEC-16-2015 00:00:00  4F2CU08112KM27774                      NaN
10587214  DEC-17-2015 00:00:00  4F2CU08112KM27774                      NaN

612 rows × 3 columns
```

Figure 6 Test Result of a Specific Vehicle

After counting the number of VIN appearing within two years, we found there are many VIN appeared hundreds of times during two years. This may be because there are some cars doing emission inspection before they are sold.

As the figure above shows, this VIN appears 612 times during these two years and there are many duplicated records at the same time.



```
                    TEST_DATE                 VIN  OVERALL_EMISSION_RESULT
 1356700  APR-22-2016 00:00:00  +1FTFX28L6VNB54156                        P
 5887815  APR-22-2016 00:00:00  +1FTFX28L6VNB54156                        P
 9899993  DEC-29-2015 00:00:00   02HGFA165X9H547013                       P
10327112  DEC-29-2015 00:00:00   02HGFA165X9H547013                       P
 9899994  DEC-08-2015 00:00:00   02HKYF18443H572372                       P
10327113  DEC-08-2015 00:00:00   02HKYF18443H572372                       P
10327114  DEC-23-2015 00:00:00             08634219                       P
 9899995  DEC-23-2015 00:00:00             08634219                       P
```

Figure 7 Vehicles with Special VINs

And there are some VINs starting with special numbers needed to be deleted and some VINs' length is not 17 digits needed to be deleted.

So, after exploring the two years small dataset and learn from it, to conclude, we need to do the following data cleaning steps for the whole 17 years data:

delete rows with any column having NaN value

according to emission type ID, only keep rows about OBD, which means 'EMISSION_TEST_TYPE_ID' equal to 'O'

choose a threshold as 2 for one month emission data, which means that only keeping VINs that appear less or equal to two times in one month

delete VIN starting with a special number

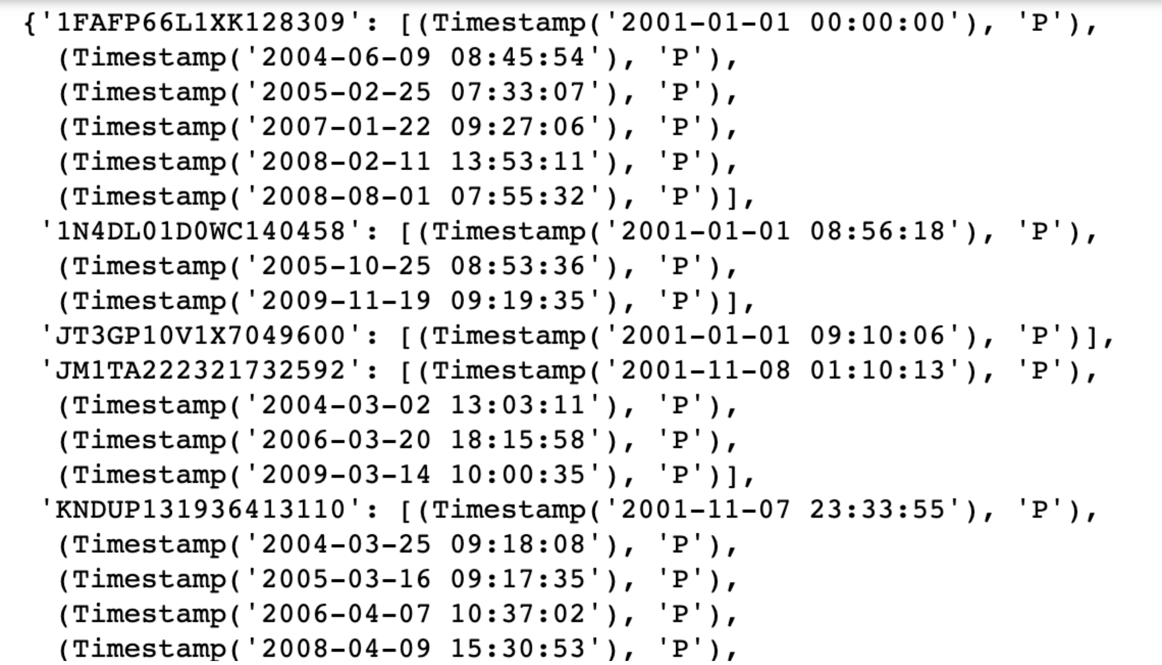
delete VINs’ length is not 12 digit
Transfer ‘‘TEST_DATE’’ format, like from ‘DEC-23-2015’ to ‘2015-12-23’
delete duplicate test date
(2) Build Dictionary

We took some detours when making the dictionary. At first we tried a slow logic to build the dictionary: read in every month(total 204 months), concat every dataframe to obtain a whole df and then do data cleaning, which is very time consuming because concat dataframe step is very slow, it needs to copy the former one, then concat the former one and the latter one.

Then we tried another logic, which is very quick and only uses one hour to help us build an intermediate dictionary.

1)read in one month dataframe
2)do data cleaning using pandas built-in data processing method
3)add the df to global dict with key as VIN

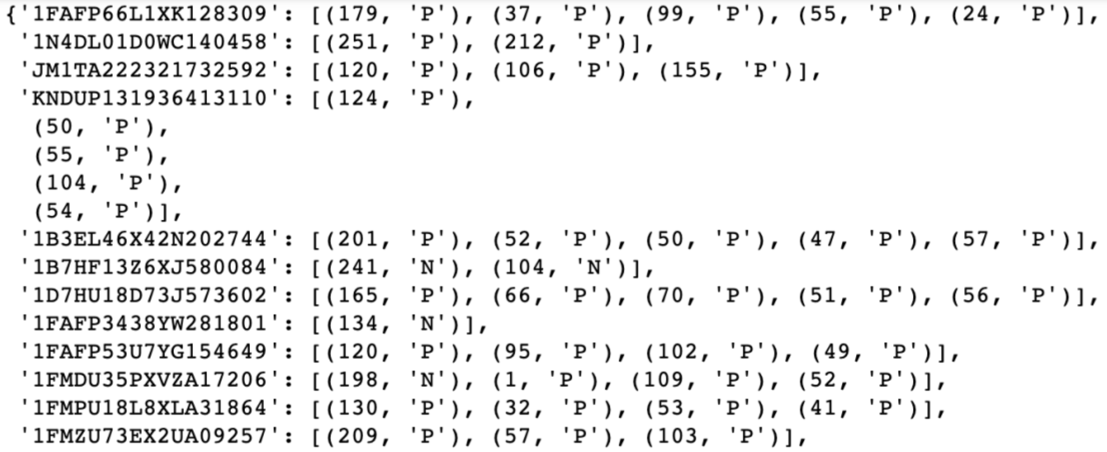
The following is our intermediate dictionary with date has been sorted, key is VIN corresponding values are test date and its test result. And the structure for the dict is:
{VIN_1: [(test_date_1, test_result_1), (test_date_2, test_result_2), (test_date_3, test_result_3)...], VIN_2: ...}



```
{'1FAFP66L1XK128309': [(Timestamp('2001-01-01 00:00:00'), 'P'),
  (Timestamp('2004-06-09 08:45:54'), 'P'),
  (Timestamp('2005-02-25 07:33:07'), 'P'),
  (Timestamp('2007-01-22 09:27:06'), 'P'),
  (Timestamp('2008-02-11 13:53:11'), 'P'),
  (Timestamp('2008-08-01 07:55:32'), 'P')],
 '1N4DL01D0WC140458': [(Timestamp('2001-01-01 08:56:18'), 'P'),
  (Timestamp('2005-10-25 08:53:36'), 'P'),
  (Timestamp('2009-11-19 09:19:35'), 'P')],
 'JT3GP10V1X7049600': [(Timestamp('2001-01-01 09:10:06'), 'P')],
 'JM1TA222321732592': [(Timestamp('2001-11-08 01:10:13'), 'P'),
  (Timestamp('2004-03-02 13:03:11'), 'P'),
  (Timestamp('2006-03-20 18:15:58'), 'P'),
  (Timestamp('2009-03-14 10:00:35'), 'P')],
 'KNDUP131936413110': [(Timestamp('2001-11-07 23:33:55'), 'P'),
  (Timestamp('2004-03-25 09:18:08'), 'P'),
  (Timestamp('2005-03-16 09:17:35'), 'P'),
  (Timestamp('2006-04-07 10:37:02'), 'P'),
  (Timestamp('2008-04-09 15:30:53'), 'P'),
```

Figure 8 Global Dict with Key as VIN

Based on the intermediate dictionary, for each VIN as key, we calculated time since last inspection as the key’s value while keeping the second test date’s result as another value for the key. The units of time since last inspection is weeks. The structure of the new dictionary is:
{VIN_1: [(time since last inspection, test_result_1), (time since last inspection, test_result_2), (time since last inspection, test_result_3)...], VIN_2: ...}



```
{'1FAFP66L1XK128309': [(179, 'P'), (37, 'P'), (99, 'P'), (55, 'P'), (24, 'P')],
 '1N4DL01D0WC140458': [(251, 'P'), (212, 'P')],
 'JM1TA222321732592': [(120, 'P'), (106, 'P'), (155, 'P')],
 'KNDUP131936413110': [(124, 'P'),
  (50, 'P'),
  (55, 'P'),
  (104, 'P'),
  (54, 'P')],
 '1B3EL46X42N202744': [(201, 'P'), (52, 'P'), (50, 'P'), (47, 'P'), (57, 'P')],
 '1B7HF13Z6XJ580084': [(241, 'N'), (104, 'N')],
 '1D7HU18D73J573602': [(165, 'P'), (66, 'P'), (70, 'P'), (51, 'P'), (56, 'P')],
 '1FAFP3438YW281801': [(134, 'N')],
 '1FAFP53U7YG154649': [(120, 'P'), (95, 'P'), (102, 'P'), (49, 'P')],
 '1FMDU35PXVZA17206': [(198, 'N'), (1, 'P'), (109, 'P'), (52, 'P')],
 '1FMPU18L8XLA31864': [(130, 'P'), (32, 'P'), (53, 'P'), (41, 'P')],
 '1FMZU73EX2UA09257': [(209, 'P'), (57, 'P'), (103, 'P')],
```

Figure 9 Global Dict with Time Intervals

(3) Data Analysis I: Failure Rate by week groups

Based on the final dictionary, we can calculate the failure rate for different ‘time since last inspections’. We have eight different time intervals: 0-10 weeks, 10-20 weeks, 20-30 weeks, 30-40 weeks, 40-50 weeks, 50-60 weeks, 60-70 weeks, 70-80 weeks and more than 80 weeks.

Based on each time interval, we calculated its corresponding failure rate. At first, we count how many data point pairs belong to this period of time, use it as the denominator; then count the number of data values that are F or N as the numerator, finally, calculate the probability of failure for each time interval. The following table shows the result:

Table 2 Failure Rate by Week Groups

Weeks	Failure Rate
0-10	0.437
10-20	0.109
20-30	0.099
30-40	0.089
40-50	0.064
50-60	0.063
60-70	0.087
70-80	0.114
>80	0.074

(4) Data Analysis II: OLS model

The OLS, Ordinary Least Square method, uses a series of predictor variables to predict the response variable (it can also be said that the response variable is regressed on the predictor variable). The best fit curve should minimize the sum of squares of the distances from each point to the straight line (i.e. the residual sum of squares, RSS for short).

Here we are trying to use OLS to detect if there is any relationship between the OBD test time interval and the OBD test result.

x: time interval since last inspection

y: test result, 0 or 1

$P(y)=0.901x+0.0006$

(5) Data Analysis III: Logistic regression

Logistic regression is a machine learning method used to solve two classification (0 or 1) problems, used to estimate the possibility of something. For example, the possibility of a user buying a certain product, the possibility of a certain patient suffering from a certain disease, and the possibility of a certain advertisement being clicked by the user, etc.

A general form of logistic regression is $\text{logit}(P) = \log(P/1+P) = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \cdots$. After applying a set of features (Xi) to calculate the probability (P) of the predicted Boolean variable is equal to 1[21].

In this part, we use one feature time since the last inspection and try to calculate the probability of the vehicle passing the emission test.

Based on the dictionary built before, if we ignore the difference of each vehicle, just use each vehicle corresponding ‘time since last inspection’ as X1, and its corresponding test result as y, when result is ‘P’, we replace ‘P’ with value 1; else, replace result with value 0. And because $y \in 0$ or 1, it is a binary classification problem, we can naturally assume that the classification result of y obeys the Bernoulli distribution and use logistic regression.

Then we can create a dataframe and apply logistic regression model on the it:

X1 (weeks)	y
50	1
55	0
...	...

After using python package sklearn and building our logistic regression model, we obtain our function: $\text{Log}(P)=0.0186x+1.388$. And the mean prediction accuracy is 0.9117, the coefficient of X1 is 0.0186, and the intercept is 1.338.

Conclusion and Discussion

(1) Failure Rate by Time Interval Group

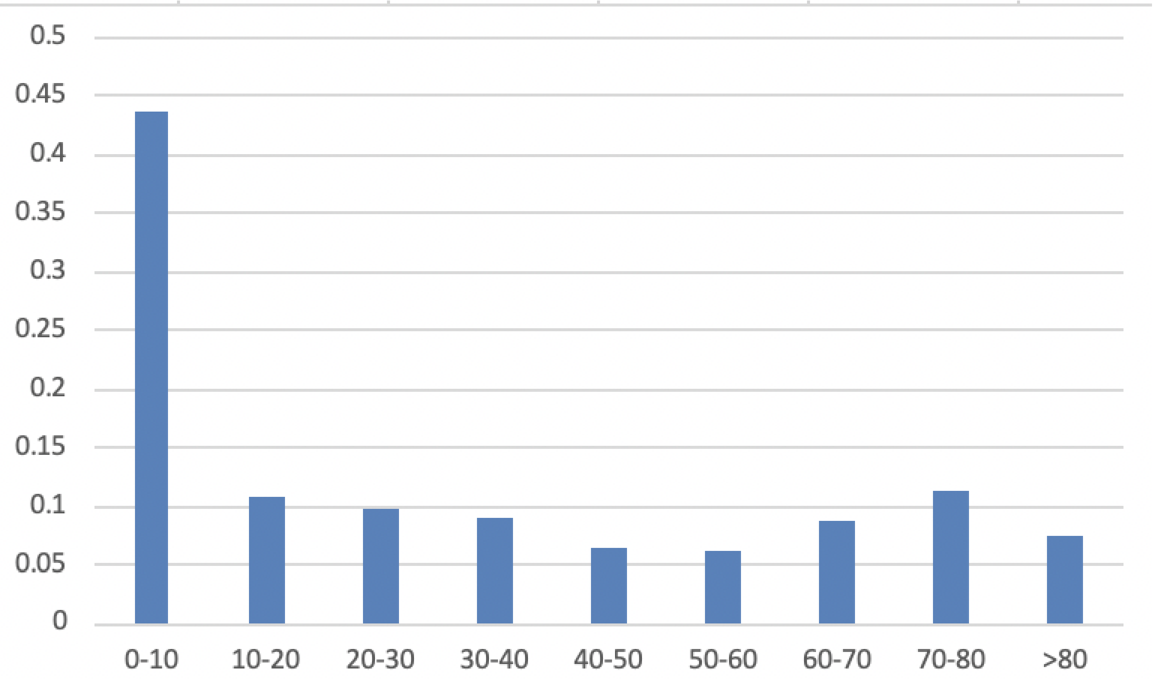


Figure 10 Failure Rate by Time Interval

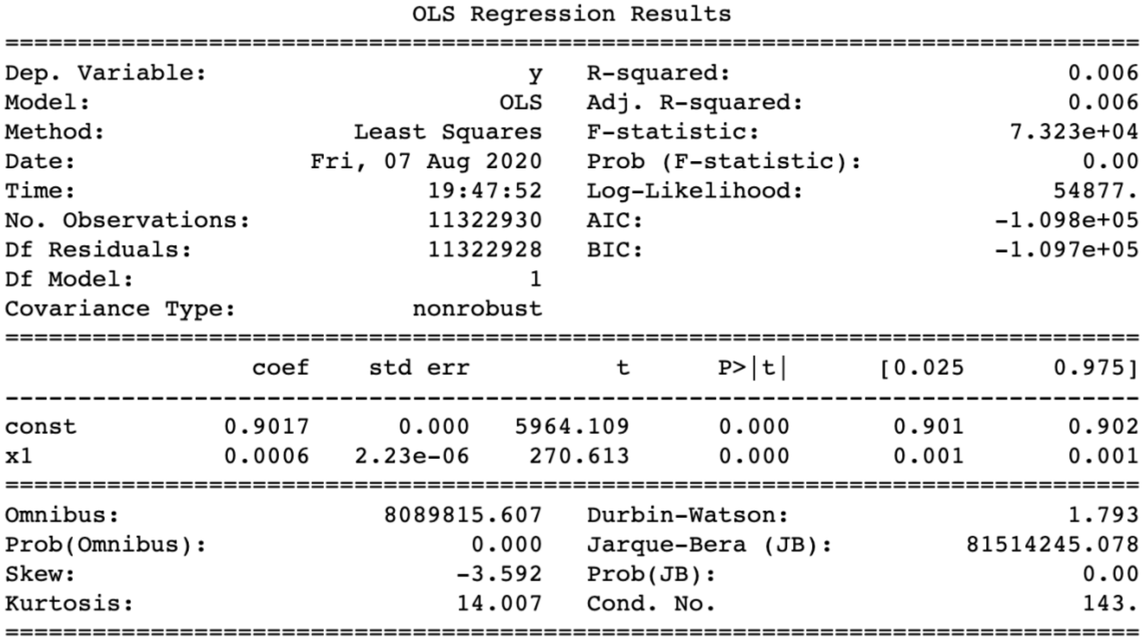
The above figure shows the trend of failure rate of different vehicles' time since the last inspection interval.

We can see from the figure that the lowest point lies in 50 weeks to 60 weeks. For our goal, to figure out whether the failure rate would be lower if we change the policy of doing emission tests once a year to half a year or so, from this once a year Pennsylvania data, we can see the lowest failure rate belongs to one year lies in the 50 weeks to 60 weeks section.

And in 0 weeks - 10 weeks, we can see failure rate is very high, this is because some cars fail the test the first time, and after repairing, then come back doing emission test again and fail it again because these cars maybe very old and already have many problems and it is hard to fix them well. And there is another situation where the consulting firm is charged with supervising whether the inspection station operates according to the rules properly. So the staff of the consulting firm would do inspection everyday to check the work of the inspection station. For example, the staff would replace the car with a broken tire in advance and then do a test to see if the test result fails. These are some reasons why the failure rate during this period of time is higher than others, and there will be cars undergoing testing in such a short period of time.

And through this picture, we can see that there is a positive u-shaped trend. This is not as we expected before: if the detection interval is shorter than one year, such as half a year, and the failure rate would drop a bit.

(2) OLS Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                 7.323e+04
Date:                Fri, 07 Aug 2020   Prob (F-statistic):               0.00
Time:                        19:47:52   Log-Likelihood:                 54877.
No. Observations:            11322930   AIC:                        -1.098e+05
Df Residuals:                11322928   BIC:                        -1.097e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.9017      0.000   5964.109      0.000       0.901       0.902
x1             0.0006   2.23e-06    270.613      0.000       0.001       0.001
==============================================================================
Omnibus:                  8089815.607   Durbin-Watson:                   1.793
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         81514245.078
Skew:                          -3.592   Prob(JB):                         0.00
Kurtosis:                      14.007   Cond. No.                         143.
==============================================================================
```

Figure 11 OLS Regression Results

From the OLS model above, we can see that there is a positive correlation between OBD test time interval and OBD test result. Since we regarded the result “Pass” as the value 1 and “Fail” as the value 0, in which case we can see if we expand the time interval of OBD testing, the probability will go closer to value 1, the result of Pass.

But in terms of coefficients, the results of linear regression model fitting are not particularly good. Because our y represents the probability of successfully passing the emission test, but the coefficient is 0.9, which is a bit large. When x increases by 1, y increases by 0.9, and x continues to increase, y may exceed 1, which is unreasonable.

(3) Logistic Regression Result analysis

The function we obtained after applying the logistic regression model is Log(P)=0.0186x+1.388. We can see the coefficient of x is 0.0186, which means that x has a positive relationship with the probability of passing emission result. The meaning of this formula about emission tests is that when we increase the interval of the car’s emission test, the probability the car will pass the test, would be bigger. In other words, the greater the interval between vehicle inspections, the more likely it is to pass the test, the failure rate would be smaller. Unlike the discussion in Conclusion 1, as the emission test interval increases, the failure rate exhibits a U-shaped change. The result of the LR model is a single positive correlation.

For the following work, we know that Colorado policy is doing emission tests twice a year. We can do the same data analysis for Colorado data, to see how the failure rate of different time intervals change and if we ignore the difference in regions, we can compare failure rate in Pennsylvania and failure rate in Colorado during the same time interval.

Another thing we can do is that, in a statistics model, we can consider adding different influencing factors, such as the model year of the car, mileage of cars, etc., to establish a more reliable statistical model, analyze the impact of various characteristics of the car on its emission test result. For example, if we use several features of vehicles as Xi to build a logistic regression model, the larger the calculated coefficient, the greater the influence of this characteristic of the coefficient on the result. In this way, we can find the main factors that affect whether the car passes the emission test. And based on these models, we can adjust the vehicle emission test policy reasonably to achieve the dual results of environmentally optimal and economically optimal.

References

[1]. Sung-hyun Baek, Jong-Wook Jang, “Implementation of integrated OBD-II connector with external network. Information Systems,” Volume 50, 2015, Pages 69-75.
[2]. Baltusis, P., "On Board Vehicle Diagnostics," SAE Technical Paper 2004-21-0009, 2004.
[3]. Le Liu, Jingyuan Li, Lihui Wang, Wei Zhao, Hongyu Qin, Yuwei Wang. 2019. “Experimental Study on Effects of OBD II Diagnostics on of Emissions for Light Vehicles.” Application of Intelligent Systems in Multi-modal Information Analytics, pages 1201-1212.
[4]. Jerzy Merkisz and Marcin Rychter, “Basic proceeding of diagnosis and strategy of decision on OBD II system,” Poznań University of Technology Institute of Internal Combustion Engines and Basic Machine Design.
[5]. H. Yun, S. Lee and O. Kwon, "Vehicle-generated data exchange protocol for Remote OBD inspection and maintenance," 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), Seogwipo, 2011, pp. 81-84.
[6]. Y. U. Devi and M. S. S. Rukmini, "IoT in connected vehicles: Challenges and issues — A review," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1864-1867.
[7]. Song S., Kim W., Kim SH., Choi G., Kim H., Youn CH, (2016) “A Prospective Cloud-Connected Vehicle Information System for C-ITS Application Services in Connected Vehicle Environment,” In: Zhang Y., Peng L., Youn CH. (eds) Cloud Computing. CloudComp 2015. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 167.
[8]. Parodi A., Maresca M., Provera M., Baglietto P, (2016) “An IoT Approach for the Connected Vehicle,” In: Mandler B. et al. (eds) Internet of Things. IoT Infrastructures. IoT360 2015. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 170.
[9]. L. Thibault, P. Pognant-Gros, P. Degeilh, K. Thanabalasingam, G. Sabiron and L. Voise, "Real-Time Air Pollution Exposure and Vehicle Emissions Estimation Using IoT, GNSS Measurements and Web-Based Simulation Models," 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 2018, pp. 1-5.
[10]. Reddy Mallidi, S Kumar; Vangipurapu, V Vineela, “IoT Based Smart Vehicle Monitoring System,” International Journal of Advanced Research in Computer Science IoT Based Smart Vehicle Monitoring System; Udaipur Vol. 9, Iss. 2, (Mar 2018): 738-741.
[11]. C. Vong, P. Wong, Z. Ma and K. Wong, "Application of RFID technology and the maximum spanning tree algorithm for solving vehicle emissions in cities on Internet of Things," 2014 IEEE World Forum on Internet of Things (WF-IoT), Seoul, 2014, pp. 347-352.
[12]. Wessel, Ryan J.. “Vehicle Emissions Inspection Programs : Equality and Impact.” (2017).
[13]. M. T. Khorshed, N. A. Sharma, K. Kumar, M. Prasad, A. B. M. S. Ali and Y. Xiang, "Integrating Internet-of-Things with the power of Cloud Computing and the intelligence of Big Data analytics — A three layered approach," 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, 2015,

pp. 1-8.

[14]. Vaidya, B., Mouftah, H.T, “IoT Applications and Services for Connected and Autonomous Electric Vehicles,” Arab J Sci Eng 45, 2559–2569 (2020).

[15]. Datta, S.K., Da Costa, R.P.F., Harri, J., Bonnet, C., 2016. “Integrating connected vehicles in Internet of Things ecosystems: Challenges and solutions”

[16]. Sharma, Sugam, Victor I. Chang, U. Sunday Tim, Johnny S. Wong and Shashi K. Gadia. “Cloud and IoT-based emerging services systems.” Cluster Computing 22 (2018): 71-91.

[17]. M. N. Iqbal et al., "Diagnostic tool and remote online diagnostic system for Euro standard vehicles," 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, 2017, pp. 415-419.

[18]. Juan R. Pimentel, “Cloud Based Vehicle Analytics: The Potential,” Oct. 02, 2015, Kettering University.

[19]. Jensen, Joshua N., "Using Consumer Accessible Mobile Devices To Collect Vehicle Emissions Compliance Data" (2017). All Regis University Theses. 841.

[20]. G. Signoretti et al., "A Dependability Evaluation for OBD-II Edge Devices: An Internet of Intelligent Vehicles Perspective," 2019 9th Latin-American Symposium on Dependable Computing (LADC), Natal, Brazil, 2019, pp. 1-9.

[21]. Acharya, Prithvi S., H. Scott Matthews, and Paul S. Fischbeck. "Data-Driven Models Support a Vision for Over-the-Air Vehicle Emission Inspections." IEEE Transactions on Intelligent Transportation Systems (2020).