Exploring the United Kingdom Anonymised Department for Transport Ministry of Transport (MOT) Anonymised Safety, Roadworthiness Test Results

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

November 4, 2021

Contents

1 Introduction **1**

2 Approach **1**

3 Analysis **2**

4 Conclusion **13**

A Miscellaneous files **14**

List of Tables

1 An unordered list of questions. 2
2 Fuel was used by type per year. 11

List of Figures

1 First use reported by year. 12

1 Introduction

The UK Ministry of Transport (MOT) is required to test cars and other light vehicles at least once a year to ensure they comply with the current road worthiness and environmental requirements. The anonymised results of these nation wide tests are made available to the public. This report details an exporation into the 2021 test results.

2 Approach

The MOT database is of interesting size with over 38 million test results. As with almost any data of this size, the steps are to locate, clean, and analyze the data.

Locate: The data is available as a ZIP file from:
`https://data.gov.uk/dataset/e3939ef8-30c7-4ca8-9c7c-ad9475cc9b2f/anonymised-mot-tes`
The ZIP file contains one CSV file for each calendar quarter. Each file has a header record, and some number of CSV lines. These are the fields in each line:
test_id, vehicle_id, test_date, test_class_id, test_type, test_result, test_mileage, postcode_area, make, model, colour, fuel_type, cylinder_capacity, and first_use_date
Details of each field are in the MOT User Guide (see Section A).

Clean: Even though the data was computer generated, one's interpretation and implementation of CSV may not be anothers. So the data lines were "cleaned" to make loading into a database seamless. This cleaning resulted in the loss of 1 line.

Analysis: The size of the database (in excess of 38 million records) made analysis using normal R and Python tools problematic. The data was loaded into a PostGres database, SQL queries were run against the database, and finally the results were analyzed using R as needed.

3 Analysis

A series of questions asked during the exploration of the MOT anonymised safety test results.

Table 1: An unordered list of questions.

#	Question	Results
1	How many raw lines are there in the MOT database? Raw lines include all headers and "bad" lines.	38,594,017
2	How many lines are there after "bad" lines are removed? (Includes all header lines.)	38,594,016
3	How many tuples are in the database?	**tuples** 38594012
4	How many unique vehicle IDs are there?	**count** 29988236

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results

test_result	count
ABA	21344
ABR	190989
ABRVE	31
F	7296397
P	28969654
PRS	2115597

5 What are the test results by type?

postcode	aba	abr	abrve	f	p	prs
AB	189	1301	0	75922	229742	9230
AL	66	932	0	27256	127169	9076
B	663	7549	4	178475	816480	75136
BA	141	1538	0	77557	252256	17708
BB	111	2146	0	54998	231491	17391
BD	206	2066	0	63318	245049	17689
BH	248	2006	0	91739	306771	22716
BL	146	1396	1	49900	205274	13060
BN	382	2347	1	107539	351605	25084
BR	74	903	0	20015	109479	9732

First 10 rows
of 119 rows.

6 What are the test results by postal code?

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results

year	count
1	1
3	1
4	3
13	1
14	1
221	1
998	1
1005	2
1010	2
1012	1

7 What are the reported first year use? There are problems with data in the database. First use years range from 1 to 2913.

First 10 rows of 145 rows.

test_id	vehicle_id
1355972487	688080473
1165432363	1278184067
222813899	453061241
1200951149	958132003
373153567	1428388413
245208109	799459405
1767117079	665896789
1425929191	64112623
1340632219	1253123325
1170038275	429391605

8 How many test IDs reported first year use are NULL?

First 10 rows of 631 rows.

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results	
		fuel_type	**used**
		CN	153
		DI	17724943
		ED	11683
		EL	93018
		FC	656
		GA	137
9	How many different types of fuel types are reported?	GB	1013
		GD	40
		HY	439059
		LN	33
		LP	13630
		OT	30484
		PE	20279130
		ST	33

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results

testedtimes	vehicles
1	21916201
2	7598108
3	418874
4	51274
5	3341
6	360
7	53
8	13
9	5
10	3
11	1
20	1
30	1
401	1

10 How many vehicles were tested how many times?

vehicle_id	testedtimes
223981155	401
950296697	30
1424313517	20
1200788799	11
554929346	10
80739093	10
750683211	10
156470364	9
1323635323	9
672614726	9
First 10 rows of 20 rows.	

11 Which were the most tested vehicles? Results limited to the 20 most tested vehicles.

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results

test_date	test_mileage	test_type	test_result
2020-02-11	94276	NT	F
2020-02-13	NA	RT	ABR
2020-02-14	NA	RT	ABR
2020-02-14	NA	RT	ABR
2020-02-17	94354	NT	P
2020-11-12	NA	RT	ABR
2020-11-12	99946	NT	F
2020-11-17	99946	RT	P
2020-11-17	NA	RT	ABR

#	Question	Results
12	What was the test history for a selected vehicle? Query limited to vehicle_id = "672614726" because previous explorations indicated this was an "interesting" vehicle.	(see above)
13	What fuel was used by type per year? (See the discussion about the range of first use years in the database.)	Table too wide to fit in this space. Information presented elsewhere (see Table 2).

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results

vehicle_id	numberinspections
223981155	379
596385553	8
1200788799	7
750683211	7
792483345	6
791792197	6
646334849	6
974650785	6
1323635323	6
459704131	6

First 10 rows
of 1,203,476 rows.

14 Which vehicles passed more than once? A vehicle could be tested and passed more than once per year.

vehicle_id	make	timestested
1200788799	LONDON TAXIS INT	11
80739093	LONDON TAXIS INT	10
750683211	LONDON TAXIS INT	10
156470364	LONDON TAXIS INT	9
1323635323	LONDON TAXIS INT	9
104372541	LONDON TAXIS INT	8
597689537	LONDON TAXIS INT	8
1215761132	LONDON TAXIS INT	7
814462287	LONDON TAXIS INT	7
291716978	LONDON TAXIS INT	7

First 10 rows
of 15,479 rows.

15 How many taxis were tested how often?

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results
16	What are the most common terms in the "make" attribute? Serves as a surrogate as to the most common make in the database. It is close, but not exact because the "make" attribute is not consistent.	**Count** / **Term** 5,653,959 FORD 4,065,830 VAUXHALL 3,420,229 VOLKSWAGEN 1,914,139 BMW 1,851,344 PEUGEOT 1,823,633 NISSAN 1,723,679 TOYOTA 1,703,307 AUDI 1,446,961 MERCEDES-BENZ 1,437,405 RENAUTL First 10 rows of 7,149 rows.
17	How old are the tested vehicles?	Approximately 50% (17,883,539) reported a first use year prior to 2011 (see Figure 1).

(Continued on the next page.)

Table 1. (Continued from the previous page.)

#	Question	Results			
		test_mileage	**first_use_date**	**make**	**model**
		999999	2003-09-15	ROVER	75
		999999	2004-11-25	ROVER	75
		999999	1999-07-03	NISSAN	ALMERA
		999999	2005-07-13	MG	ZT
		999999	2007-09-03	MERCEDES	SPRINTER
18	Which vehicles have the most miles? (Not truely answerable because it appears the database uses 999999 as some sort of indicator.)	999999	2000-09-11	PEUGEOT	406
		999999	2004-05-29	ROVER	75
		999999	2001-03-01	AUDI	TT
		999999	2004-03-26	ROVER	75
		999999	2003-05-02	ROVER	75
		First 10 of 37,586,720 rows.			

(Last page.)

Plotting the cumulative reported first use values (see Figure 1) raises hints and ideas about aspects that are not captured in the MOT database. These include:

1. Excepting the period prior to 1900 and from 2018 onward, the graph is almost linear on a logarithmic scale, implying an exponential rate of change. But what is that rate?
2. Is there a correlation between number of cars still in use, and periods of recession or depression?
3. Do the "flat" periods on the plot correlate to changes in population, personal wealth or income?
4. How has the curve changed over time? If we were to look at the data from 5 or 10 years ago could we estimate how often cars are "retired from service" or how many are added to service?
5. Has the effective cost of a vehicle contributed to more vehicles being put into service? How much would the single car from 1854 cost compared to the "average" car from 2018?

Table 2: Fuel was used by type per year. See the discussion about the range of first use years in the database.

year	cn	di	ed	el	fc	ga	gb	gd	hy	ln	lp	ot	pe	st
1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
3	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	NA
13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
14	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
221	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
998	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1005	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1010	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	1	NA
1012	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1013	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1014	NA	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA
1015	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	NA
1016	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	NA
1017	NA	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	NA
1019	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1087	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1197	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1199	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1212	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA

First 20 rows
of 146 rows.

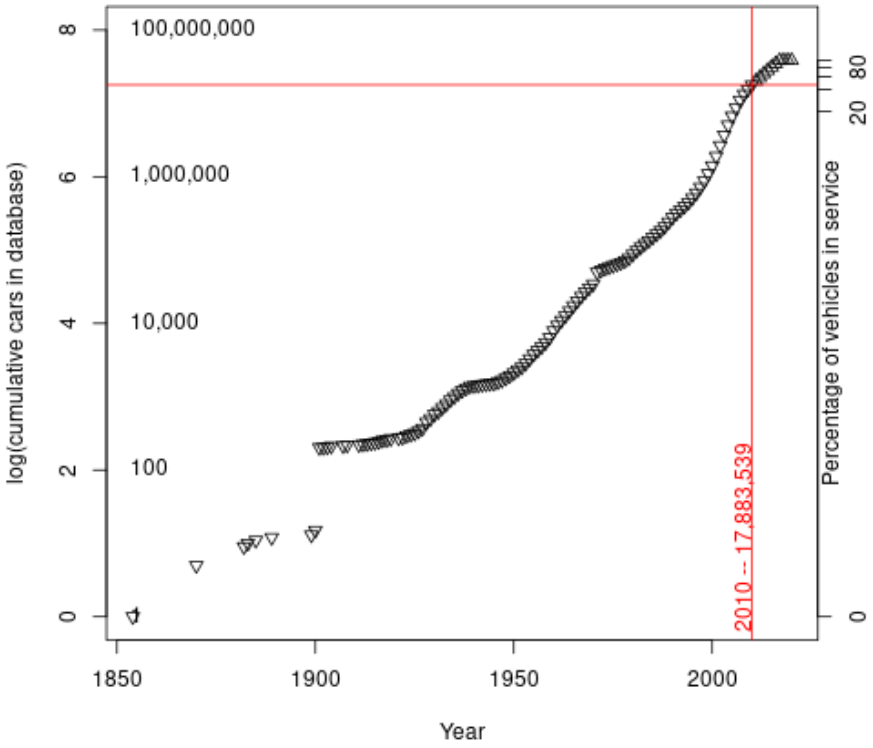

Figure 1: First use reported by year. The range of first use year reports is wide. From 1 in 1854 to 2,792,959 in 2017. The number of first year uses for the period 2018 – 2020 are considerably lower. Vehicles have to be at least 3 years old before their first test. The vertical red line shows that by 2010 there were 13,248,945 vehicles reportedly first used. The horizontal red line shows that the approximately 50% mark for the entire time period. The horizontal red line does not appear in the middle of the plot because the Y-axis is a logarithmic scale to encompass the wide range of values. The values inside the plotting area on the left hand-side are the "normal" values that correspond to the exponents.

4 Conclusion

We explored the UK's Ministry of Transport publicly available database of vehicle test results. The database contains over 38 million individual test results for almost 30 million unique vehicles from as long ago as 1854. Exploring the database identified a variety of different fuel types in use, errors when entering first use data, a number of vehicles that were inspected numerous times within a few days, and some vehicle types there were inspected more than others. The R source code and SQL commands used to create, populate, and query the database are included in this report.

A Miscellaneous files

A collection of files used in the creation of this report.

- fileDownload.sh – download, unzip, and "clean" MOT data before loading into the database
- createTablesMOT.sql – SQL commands to create MOT database
- populateTablesMOT.sql – postgres commands to populate the MOT database (based on other files)
- queryTablesMOT.sql – SQL commands to test that MOT database was loaded correctly
- exploreMOTdata.R – R script to explore the MOT database
- termFrequency.sh – A bash shell script tailored to report how many times unique terms were used in the "make" attribute of the mot relation.
- MOT_user_guide_v4.docx – MOT User's Guide

The embedded files can be extracted using an Adobe reader tool. The files may not be extractable using a web browser.