

Big Data

6CS030

2020

Contents

Introduction to Big Data.....	3
Introduction to the dataset chosen	4
Importing of the data.....	4
Analysis of data.....	6
Visualisation of data.....	8
Car Test Revenue 2018	8
Number of female HGV tests over time.....	8
Advantages & Disadvantages.....	9
Conclusion and way forward.....	10
Appendix.....	11
Data preparation and cleansing.....	11
Dataset 1.....	11
Dataset 2	12
Code used for queries	13
Query 1	13
Query 2	13
Query 3	13
References	14
Bibliography	14

Introduction to Big Data

Big Data as a conceptual idea is not a recent thing. The accumulation of large data sets has occurred, with regularity, across history. One of the most well-known types would be in the form of a census, the first known to have occurred almost 6000 years ago (Population Reference Bureau, 2019).

Present day/modern Big Data is dynamically evolving and changing as the world is becoming more and more interconnected. A buzz term of recent years “The internet of things” (Kevin Ashton, 2009) is often used to encapsulate the addition of interconnected technology with everyday items, such as smart washing machines (Siemens, 2020). Each one of these ‘connected’ devices through simple necessity of function also now creates data in some form. The accumulation of this generated data is at the core of Big Data. When describing Big Data, it is often characterised using words starting with the letter V. Originally there were three.

- Volume, referring to the sheer amount of data being generated.
- Velocity used to describe how swiftly the data changes.
- Variety refers to the many different data types.

Though over time companies and organisations have attempted to add additional descriptors with varying success. Arguably the most accepted and used would be Veracity and its definition of the trustworthiness of the data.

If one side of the Big Data coin is data generation, then the other is arguably, data analysis. Analysis could be described as the process of converting raw data into value. This ‘value’ could take many forms from increased productivity in a warehouse, or to generating significant strategic business value for a telecoms company (O,Yazidi Alaoui et al. 2019).

Whilst in the main, Big Data and its analysis is generally seen as a societal positive, consideration should be given to the possible social, ethical and legal implications associated with the activity. Some of these negatives were brought to light in 2013 with the Global surveillance disclosures made public by the whistle blower Edward Snowden (M, Gidda, 2013).

Introduction to the dataset chosen

The first dataset that will be used has been obtained from the Department for Transport (DFT). It is a list of all Large Goods Vehicle (LGV) pass rates by gender and month, covering the period April 2007 through to September 2019. The original file name is “DRT0501” and is provided in Open Document Spreadsheet (ODS) file format. It is available from <https://www.gov.uk/government/statistical-data-sets/driving-test-statistics-drt>. The dataset DRT0501 will be referred to as “dataset 1” when referenced in the remainder of this report.

The second dataset used is also produced by the DFT and also available at the same webpage. This dataset is similar to the first, except that this covers car tests, covering the same time period. The original file name is “DRT0201” and is provided in Open Document Spreadsheet (ODS) file format. The dataset DRT0503 will be referred to as “dataset 2” when referenced in the remainder of this report.

Both of the datasets offer insight into UK car and LGV test rates across a number of years with additional breakdown by gender as well as year and month. Analysis of this information could lead to interesting statistics that could help influence many automotive associated areas such as car sales or car insurance.

Many efforts have been made to find and use a suitable unstructured/semi structured dataset that would, both be practical and link to the structured datasets already selected. However, it has not proven possible on this occasion. Regardless, it is still possible to discuss the merits and commonalities that exist in such datasets through the lens of Big Data and more specifically what would be expected in terms of this report. Typically, any unstructured or semi structured dataset has a more flexible approach to how data is stored over the traditional structured datasets.

Importing of the data

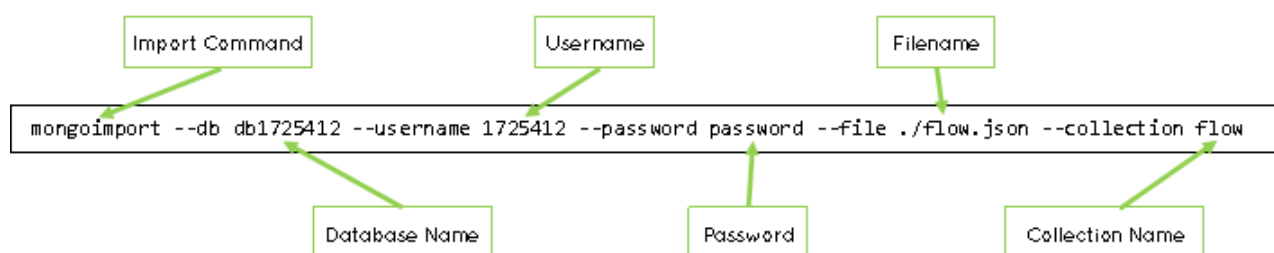
Both Dataset 1 and Dataset 2 were imported into Oracle using the SQL Developer software. Oracle was chosen as the two datasets are of a structured format and Oracle was deemed to be much more suited over MongoDB and HadHoop for handling that type of data. One key reason, from a usability perspective is the GUI available with Oracle. Before the datasets could be imported a number of steps were taken to ‘clean/cleanse’ the data. This involved removal of non-relevant data at the top and bottom of each file as well as changing the format of certain columns, most notably the ‘date’ columns into a format that would be more usable with SQL Developer. With the datasets used there were no gaps in the data where any totals had to be generated using techniques such as averaging

neighbouring data. Also, there were no apparent 'rouge' data totals that stood out against the flow of data within the tables.

Both datasets were loaded into their own respective tables and given relevant names for the tables themselves as well as suitable column headings. In total each dataset required around twenty-five steps of preparation and cleansing before being imported. Full details of all steps taken for each of the datasets can be found in the appendix.

Following on from the final paragraph, in the Introduction to the Dataset Chosen section of this report, the necessary steps needed to clean and prepare unstructured data will now be discussed.

When looking at the MongoDB software and how data is imported into the DBMS. MongoDB allows for a single command line to import an entire dataset easily. You can see below how the command is structured. When working from the command line, in this way it is usual that you will either get success, i.e. a confirmation will appear on the console or alternatively you will get an error message. The importation process is very much, an all or nothing equation. As to cleansing of the files themselves, usually this is not required as they come in a ready to use format, this is in a large part due to the heavy syntax applied throughout the documents.



Example of a MongoDB import command

Analysis of data

The first query that was ran on dataset 1 was to total the number of tests by year and also show the number of tests as both a percentage and numerical figure for each gender and overall total. The query was written in SQL and used the SUM function as well as some other mathematical elements. The code used can be found in the appendix labelled Query 1.

The query produced the following results:

	YEAR	Male Tests	Male %	Female Tests	Female %	TOTAL
1	2007	50387	93.91	3257	6.07	53656
2	2008	65222	94	4151	5.98	69386
3	2009	47111	93.06	3512	6.94	50626
4	2010	38397	93.26	2777	6.74	41174
5	2011	43984	93.45	3085	6.55	47069
6	2012	43565	93.2	3179	6.8	46744
7	2013	44166	92.96	3345	7.04	47511
8	2014	47932	93.02	3598	6.98	51530
9	2015	62422	92.96	4727	7.04	67149
10	2016	72044	92.81	5579	7.19	77623
11	2017	66735	92.12	5711	7.88	72446
12	2018	66571	91.46	6216	8.54	72787
13	2019	50695	90.73	5177	9.27	55872

*Results from query 1, showing the total number and proportion of HGV tests by gender and year.
n.b. 2019 data is not a complete year and only covers nine months*

Given that the above data relates specifically to Large Goods Vehicle tests, it is clear to see that there is a large ratio of male to female candidates. Though it is worth noting that there has been an increase of almost 50% over the date range covered within the dataset. It is also worth noting that the data for 2019 is not representative of a complete year and only contains nine months of figures. The decision was made to leave the incomplete data in the table, for illustrative purposes.

A similar query was then run on dataset 2, to show the same information, but this time based on car tests and not HGV tests. The code used can be found in the appendix labelled Query 2.

The query produced the following results:

	YEAR	Male Tests	Male %	Female Tests	Female %	TOTAL
1	2007	643195	49.05	667849	50.93	1311349
2	2008	871594	48.84	912804	51.14	1784752
3	2009	779964	49.05	809910	50.94	1590045
4	2010	731921	48.36	781626	51.64	1513561
5	2011	769054	47.64	845263	52.36	1614323
6	2012	713380	47.47	789551	52.53	1502942
7	2013	677123	47.73	741626	52.27	1418756
8	2014	724570	47.91	787911	52.09	1512484
9	2015	742492	47.89	808066	52.11	1550559
10	2016	787795	47.8	860152	52.2	1647955
11	2017	828227	46.83	940307	53.17	1768539
12	2018	794861	46.96	897920	53.04	1692782
13	2019	567571	47.27	633006	52.73	1200577

Results from query 2 showing the total number and proportion of car tests by gender and year.

n.b. 2019 data is not a complete year and only covers nine months

Analysis of the car test data reveals that every year, there are marginally more female tests undertaken than male tests. Also, there is no clear pattern in the total tests undertaken from year to year, neither an upward nor downward trend is apparent.

A further query was performed against dataset 2 to estimate the revenue generated from car tests around the UK. The current price of a car test is £62, and this figure was used to make the calculations. The code used can be found in the appendix labelled Query 3.

	YEAR	Car Revenue
1	2007	£72,124,195
2	2008	£98,161,360
3	2009	£87,452,475
4	2010	£83,245,855
5	2011	£88,787,765
6	2012	£82,661,810
7	2013	£78,031,580
8	2014	£83,186,620
9	2015	£85,280,745
10	2016	£90,637,525
11	2017	£97,269,645
12	2018	£93,103,010
13	2019	£66,031,735

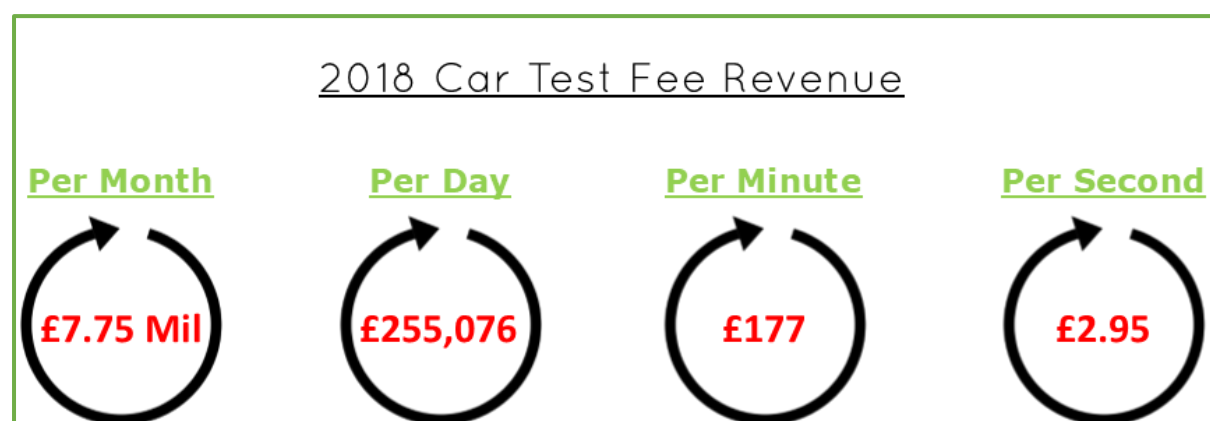
Results from query 3 showing the estimated revenue generated from car tests within the UK

n.b. 2019 data is not a complete year and only covers nine months

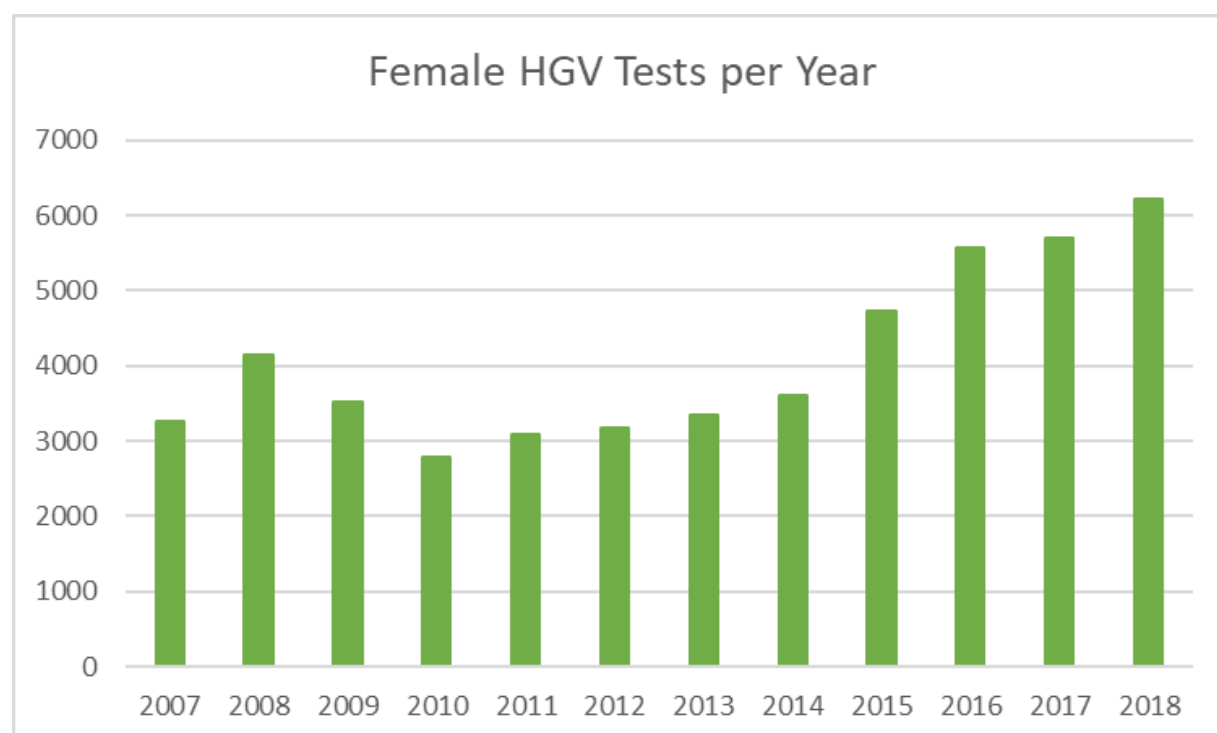
Visualisation of data

After performing the analysis upon both datasets, the following visualisations were created using MS Excel & MS Publisher. The chosen format is that of simple eye-catching statistics that are representative of the analysis results.




Car Test Revenue 2018



Number of female HGV tests over time



Advantages & Disadvantages

DBMS	Advantages	Disadvantages
	<p>Point: Simplification of table creation and data importation.</p> <p>Justification: Using the Oracle SQL developer tool. The software allows the user to simply load a <u>suitable</u> file, either csv or equivalent and then after some intuitive UI screens a table is generated and ready to be queried.</p>	<p>Point: Data cleansing preparation not being fully documented.</p> <p>Justification: If the steps taken to prepare data to be imported are not correctly documented then should anyone desire to replicate the results or further analysis applied to the data, this would prove difficult. Another possible issue that exists around data cleansing is the introduction of a bias where missing data elements existed, and new data is generated.</p>
	<p>Point: Ability to use unstructured data.</p> <p>Justification: An advantage to using this approach for handling Big Data is the flexibility of being able to use unstructured data. Being able to swiftly and efficiently reshape the database to accommodate datasets with differing numbers of key value pairs allows for rapid analysis and results.</p>	<p>Point: Inability to see a document clearly within a collection.</p> <p>Justification: When trying to view a document within a collection it can become more difficult depending on the size of the document and the number of key value pairs that it holds, as well as the nesting of those elements. This all adds up to a very nonuser friendly output, on the screen for the user. It is worth noting that Python makes this better, but only marginally.</p>
	<p>Point: Scalability to handle vast amounts of data.</p> <p>Justification: Researching the capabilities of the software, it is apparent that Hadoop is capable, via means of distributed storage and processing, of allowing an entity to run applications across thousands of nodes, processing huge amounts of data (thousands of terabytes).</p>	<p>Point: Lack of s GUI</p> <p>Justification: The Hadoop suite of applications does not have a GUI (graphical User Interface). A GUI can have a number of advantages over a traditional console-based interface. With console-based interfaces like Hadoop, there is a greater possibility in user error when typing commands. This is something that would be reduced with a GUI as typically the user will simply click a pre created command icon or menu item and be able to complete a complex action.</p>

Conclusion and way forward

Big data is something that is only going to get bigger. As more and more individuals generate increasingly large amounts of data then as to will the demand for big data analysis increase. In this report three DBMS solutions are looked at, as well as exploring both structured and unstructured data. Both Hadoop and MongoDB lend themselves to processing and handling unstructured data and Oracle lends itself moreover to structured data sources, though it is worth noting that newer versions of Oracle are becoming more capable in handling unstructured data.

Whilst data has typically been in a structured format, financial records, staff databases, medical records etc. these will not go away nor change. It is, however, foreseeable that the majority of new big data sets will be in an unstructured format. This could indicate that the processing of this data may be better suited to Hadoop, due to its ability to scale up and distribute its analysis amongst many nodes (individual computers).

Appendix

Data preparation and cleansing

Please note that where row, column or cell references are used, they are done so based upon all previous amendments and deletions. For example, dataset 1 in its original pre-cleansed state contains footer style information on rows 174-180, as well as header information on rows 1-9. When recording the steps taken to cleanse the data it will be recorded as:

- Deletion of rows 1-9
- Deletion of rows 165-171

Dataset 1

In preparing to import dataset 1 into SQL, the following changes were made in Microsoft Excel:

- Deletion of rows 1-7
- Deletion of row 2
- Deletion of rows 166-172
- Changed all percentage columns to two decimal places
- Renamed cell A1 to Month
- Renamed cell B1 to Male_Tests
- Renamed cell C1 to Male_Test_Passes
- Renamed cell D1 to Male_Pass_Percentage
- Renamed cell E1 to Female_Tests
- Renamed cell F1 to Female_Test_Passes
- Renamed cell G1 to Female_Pass_Percentage
- Renamed cell H1 to Total_Tests
- Renamed cell I1 to Total_Test_Passes
- Renamed cell J1 to Total_Pass_Percentage
- All row heights set to the same height for ease of viewing
- All borders removed
- All cell fill removed
- Deletion of columns B,F & J
- Shifted all columns one to the right
- Split the date into year into column A and the month into column B
- Renamed cell A1 to Year
- Saved as .xlsx file types

The file was then imported into Oracle using SQL Developer making the following choices:

- Table name set to LGVPass_Rate_By_Month
- Set Male_Tests, Male_Test_Passes to size/precision 10 and scale 0
- Set Female_Tests, Female_Test_Passes to size/precision 10 and scale 0
- Set Total_Tests, Total_Test_Passes to size/precision 10 and scale 0
- Set all Percentage columns to size/precision 4 and scale 2
- Set Year to size/precision 5

Dataset 2

In preparing to import dataset 2 into SQL, the following changes were made in Microsoft Excel:

- Deletion of rows 1-7
- Deletion of rows 152-172
- Deletion of columns F & J
- Renamed cell A1 to Year
- Renamed cell B1 to Month
- Renamed cell C1 to Male_Tests
- Renamed cell D1 to Male_Test_Passes
- Renamed cell E1 to Male_Pass_Percentage
- Renamed cell F1 to Female_Tests
- Renamed cell G1 to Female_Test_Passes
- Renamed cell H1 to Female_Pass_Percentage
- Renamed cell I1 to Total_Tests
- Renamed cell J1 to Total_Test_Passes
- Renamed cell K1 to Total_Pass_Percentage
- Split date into year and month in columns A&B
- All row heights set to the same height for ease of viewing
- All borders removed
- All cell fill removed
- Saved as .xlsx file types

The file was then imported into Oracle using SQL Developer making the following choices:

- Table name set to CARPass_Rate_By_Month
- Set Year to size/precision 5
- Set Male_Tests, Male_Test_Passes to size/precision 10 and scale 0
- Set Female_Tests, Female_Test_Passes to size/precision 10 and scale 0
- Set Total_Tests, Total_Test_Passes to size/precision 10 and scale 0
- Set all Percentage columns to size/precision 4 and scale 2

Appendix that documents all the code used for the practical work.

Code used for queries

Query 1

```
SELECT
YEAR,
Sum( MALE_TESTS) as "Male Tests",
(Round(Sum (Male_Tests)/Sum(Total_Tests),4)*100) as "Male %",
Sum( FEMALE_TESTS) as "Female Tests",
(Round(Sum (female_Tests)/Sum(Total_Tests),4)*100) as "Female %",
Sum( TOTAL_TESTS) as Total
```

Query 2

```
SELECT
YEAR,
Sum( MALE_TESTS) as "Male Tests",
(Round(Sum (Male_Tests)/Sum(Total_Tests),4)*100) as "Male %",
Sum( FEMALE_TESTS) as "Female Tests",
(Round(Sum (female_Tests)/Sum(Total_Tests),4)*100) as "Female %",
Sum( TOTAL_TESTS) as Total
```

Query 3

```
select
car.year,
TO_CHAR(Sum((car.Total_Tests)*55), 'L999,999,999') as "Car Revenue"
from
Carpass_rate_by_month car
```

References

Population Reference Bureau (2019) Milestones and Moments in Global Census History [Online]. [Accessed 03-02-20]. Available from: <https://www.prb.org/milestones-global-census-history/>

Ashton, K. (2009) RFID Journal : That 'Internet of Things' Thing [Online]. [Accessed 03-02-20]. Available from: <https://www.rfidjournal.com/articles/view?4986>

Siemens (2020) Our Smart Washing Machines and Dryers [Online]. [Accessed 03-02-20]. Available from: <https://www.siemens-home.bsh-group.com/uk/inspiration/innovation/connected>

O. Yazidi Alaoui et al. (2019) CREATING STRATEGIC BUSINESS VALUE FROM BIG DATA ANALYSIS – APPLICATION TELECOM NETWORK DATA AND PLANNING DOCUMENTS. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. [Online] XLII-4-W16691-695. [online]. Available from: <https://doi.org/article/64231055f1804a45874bb16930dfa176>.

M Gidda (2013). The Guardian: Edward Snowden and the NSA files – Timeline [Online]. [Accessed 03-02-20]. Available from: <https://www.theguardian.com/world/2013/jun/23/edward-snowden-nsa-files-timeline>

Bibliography

Ranjan, J. (2019) The 10 Vs of Big Data framework in the Context of 5 Industry Verticals. Productivity. [Online] 59 (4), 324–342. [online]. Available from: <http://search.proquest.com/docview/2197772407/>

Owens Jonathan R et al. (2013) 'Big Data Analysis', in Hadoop Real-World Solutions Cookbook. Packt Publishing. pp. 1–1. [online]. Available from: <https://app.knovel.com/hotlink/pdf/rcid:kpHRWSC003/id:kt00BEGW81/hadoop-real-world-solutions/big-data-analysis?kpromoter=Summon>.