CMR
INSTITUTE OF TECHNOLOGY

USN										

Internal Assessment Test I – May 2024

Sub:	**Big Data Analytics**							**Sub Code:**	**22MCA412**
Date:	**25/05/2024**	**Duration:**	**90 mins**	**Max Marks:**	**50**	**Sem:**	**IV**	**Branch:**	**MCA**

Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

	PART I	MARKS	OBE	
			CO	RBT
1	Explain about the different data sources in big data. **OR**	2+8	CO3	L4
2	Write a note on outliers and outlier detection in big data analytics	4+6	CO4	L4
3	**PART II** Write a note on applications of big data analytics. **OR**	10	CO4	L4
4	Explain the types of data elements in big data.	10	CO2	L2

5	**PART III** Discuss about categorization in big data. **OR**	10	CO2	L2
6	Write a note on crowd sourcing in big data.	10	CO2	L3
7	**PART IV** Explain the following in big data i) Data Discovery ii) Open source technology for BDA **OR**	5+5	CO3	L3
8	Explain the V's of big data analytics	2+8	CO3	L3
9	**PART V** Discuss about the steps to solve a problem using z scores and solve the following: **1257,1218,1158,5130,3630,21467,1274,4582,6374,2251,2623** **OR**	10	CO4	L4
10	Discuss about the steps to solve a problem using box plot and solve the following: 99;56;78;55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90	10	CO4	L4

1.Types of Data Sources

Data can originate from a variety of different sources. They are as follows:

- Transactional data
- Unstructured data
- Qualitative/Expert based data
- Data poolers
- Publicly available data

Transactional Data: Transactions are the first important source of data. Transactional data consist of structured, low level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim, cash transfer, credit card payment). This type of data is usually stored in massive online transaction processing (OLTP) relational databases. It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.

Unstructured data: Embedded in text documents (e.g., emails, web pages, claim forms) or multimedia content can also be interesting to analyze. However, these sources typically require extensive pre-processing before they can be successfully included in an analytical exercise.

Qualitative/Expert based data: Another important source of data is qualitative, expert based data. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager). The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run. This will steer the modeling in the right direction and allow you to interpret the analytical results from the right perspective. A popular example of applying expert based validation is checking the univariate signs of a regression model. For example, one would expect a priori that higher debt has an adverse.

Data poolers: Nowadays, data poolers are becoming more and more important in the industry. Popular examples are Dun & Bradstreet, Bureau Van Dijck, and Thomson Reuters. The core business of these companies is to gather data in a particular setting (e.g., credit risk, marketing), build models with it, and sell the output of these models (e.g., scores), possibly together with the underlying raw data, to interested customers. A popular example of this in the United States is the FICO score, which is a credit score ranging between 300 and 850 that is provided by the three most important credit bureaus: Experian, Equifax, and TransUnion. Many financial institutions use these FICO scores either as their final internal model or as a benchmark against an internally developed credit scorecard.

Publicly available data: Finally, plenty of publicly available data can be included in the analytical exercise. A first important example is macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on. By including this type of data in an analytical model, it will become possible to see how the model varies with the state of the economy. This is especially relevant in a credit risk setting, where typically all models need to be thoroughly stress tested. In addition, social media data from Facebook, Twitter, and others can be an important source of information. However, one needs to be careful here and make sure that all data gathering respects both local and international privacy regulations.

2. Outliers and outlier detection

Outliers are extreme observations that are very dissimilar to the rest of the population. Actually, two types of outliers are:

- Valid observations (e.g salary of boss is $1 million)
- Invalid observations (e.g age is 300 years)

Both are univariate outliers in the sense that they are outlying on one dimension, outliers can be hidden in unidimensional views of the data.

Multivariate outliers are observations that are outlying in multiple dimensions. Two important steps in dealing with outliers are detection and treatment

A first check for outlier is to calculate the minimum and maximum values for each data element

Next various graphical tools can be used to detect outliers.

Example:

- Histograms
- Box plots
- Z-scores

Figure 1.4 presents an example of a distribution for age whereby the circled areas clearly represents outlier.

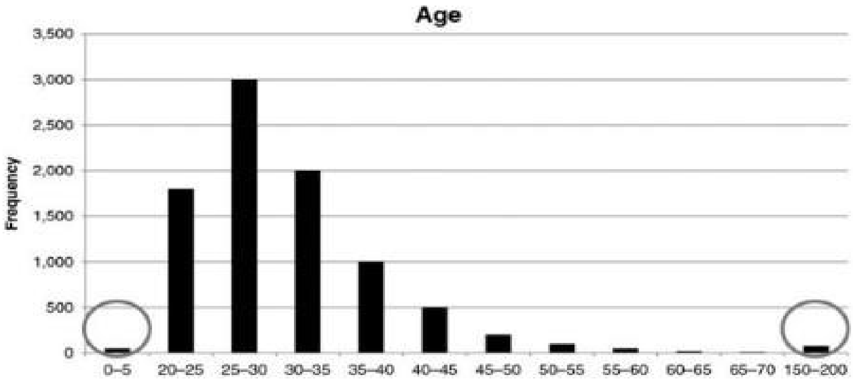


Figure 1.4 Histogram representations of outliers

Another visual mechanism is **box plots**. A box plot represents three key qualifiers of data. They are:

- First Quartile (25% of the observations have a lower value)
- Median (50% of the observations have a lower value)
- Third Quartile (75% of the observations are lower value)

All three quartiles are represented as a box. The minimum and maximum values are then also added unless they are too far from the edges of the box. Too far away is then 1.5*Interquartile Range(IQR=Q3- Q1). The figure 1.5 gives an example of a box plot in which the three outliers can be seen.

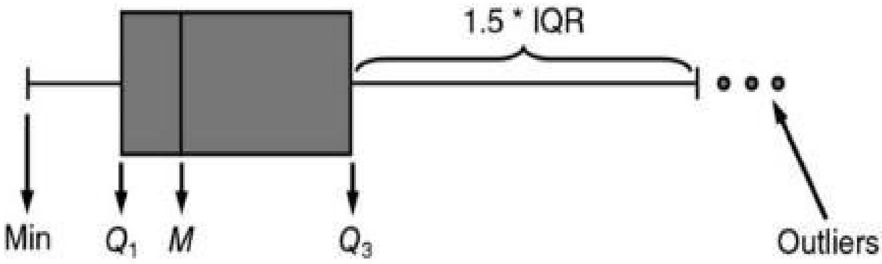


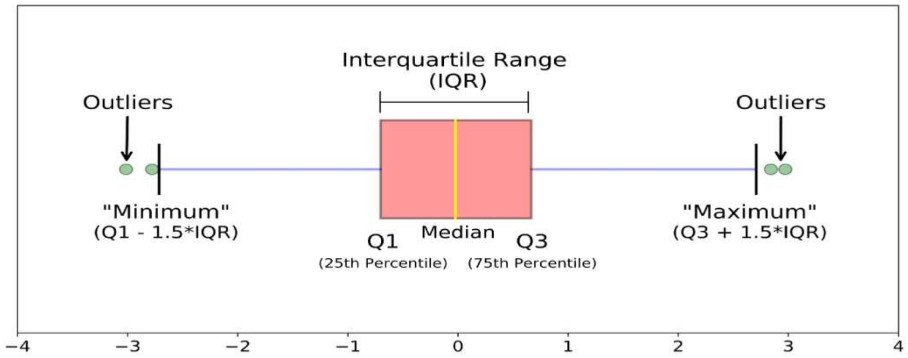
Figure 1.5: Box Plots for Outlier Detection

Box-Plot:

What is a box and whisker plot?

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.



Z-Scores: Another way to identify outliers is to calculate **z-scores**, measuring how many standard deviations an observation lies away from the mean as follows:

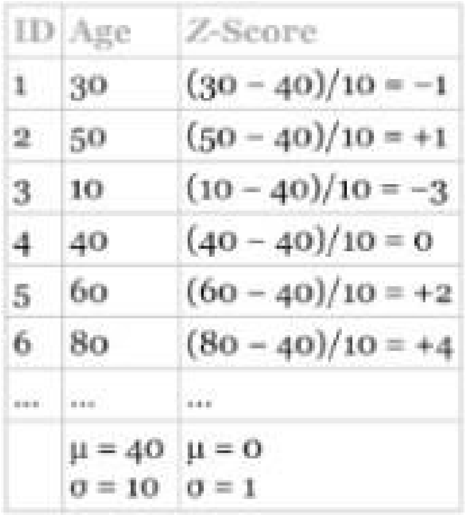
$$z_i = \frac{x_i - \mu}{\sigma}$$

Where, μ-represents the average of the variables σ-

represents standard deviation

an example is given in table 1.3. Note that by definition, the z-scores will have 0-mean and unit standard deviation.

Table 1.3: z-scores for Outlier Detection



ID	Age	Z-Score
1	30	(30 – 40)/10 = –1
2	50	(50 – 40)/10 = +1
3	10	(10 – 40)/10 = –3
4	40	(40 – 40)/10 = 0
5	60	(60 – 40)/10 = +2
6	80	(80 – 40)/10 = +4
...	...	...

	μ = 40
	σ = 10
		μ = 0
		σ = 1

A practical rule of thumb then defines outliers when the absolute value of the z-score |z| is bigger than 3.

Note that the z-scores relies on the normal distribution.

Multivariate outliers can be detected by fitting regression lies and inspecting the observation with large errors. (e.g residual plot). Alternative methods are *clustering*. Some analytical techniques like *decision trees*, *neural networks* are fairly robust with respect to outliers. Various schemes exist to deal with outliers. It highly depends on whether the outlier represents a *valid* or invalid *observation*.

A popular scheme for is truncation/capping/winsorising. One here by imposes lower limit and upper limit on a variable and any values below/above are brought back to the limits. The limits can be calculated using the z-scores or the IQR (which is more robust than the z-scores)

Calculating upper and lower limit using z-score is: Upper

limit=μ+3σ

Lower limit= μ-3σ

This calculation is shown in the figure 1.6.

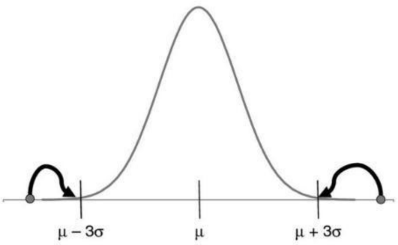


Figure 1.6: Using the z-scores for Truncation

Another way of calculating the upper and lower limit is using box plot IQR. Upper

limit=M+3s

Lower limit=M-3s

Where, M is Median and s=IQR/(2*0.6745)[3]

3. Applications of big data

Example Applications

Analytics is everywhere and strongly embedded in our daily lives.

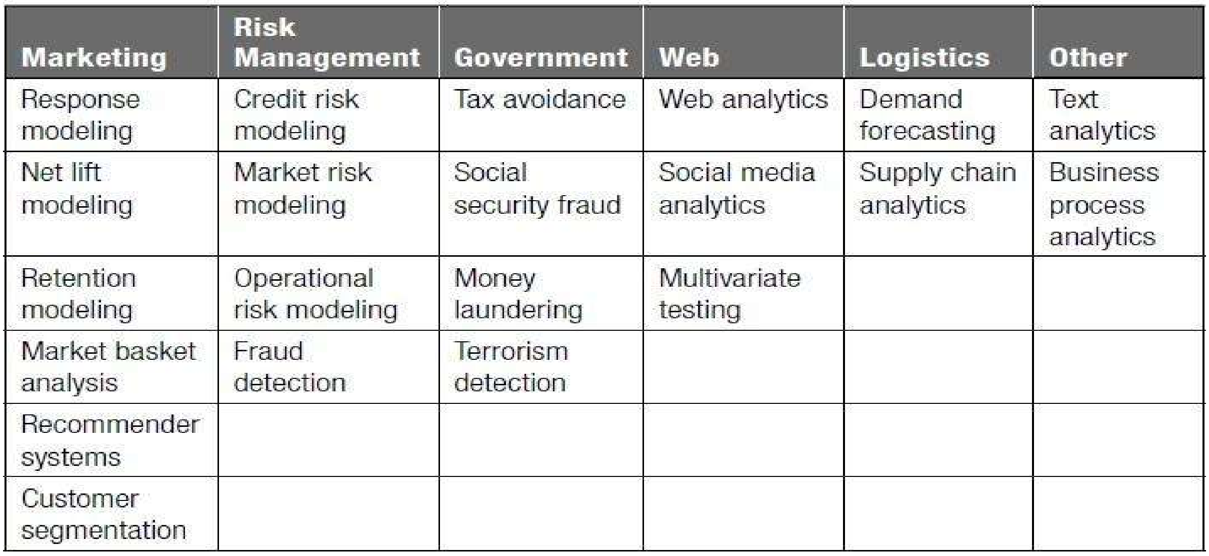
The relevance, importance and impact of analytics are now bigger than ever before and, given that more and more data are being collected and that there is strategic value in knowing what is hidden in data, analytics will continue to grow.

Physical mail box: a catalogue sent to us through mail most probably as a result of a response modeling analytical exercise that indicated, given my characteristics and previous purchase behavior, we are likely to buy one or more products from it.

Behavioral Scoring Model: Checking account balance of the customer from the past 12 months and credit payments during that period, together with other kinds of information available to the bank, to predict whether a customer will default on the loan during the next year.

Social Media: As we logged on to my Facebook page, the social ads appearing there were based on analyzing all information (posts, pictures, my friends and their behavior, etc.) available to Facebook. Twitter posts will be analyzed (possibly in real time) by social media analytics to understand both the subject tweets and the sentiment of them.

Table 1.1 Example Analytics Applications



Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

4. types of data elements

It is important to appropriately consider the different types of data elements at the start of the analysis. The different types of data elements can be considered:

- Continuous
- Categorical

Continuous: These are data elements that are defined on an interval that can be limited or unlimited. Examples include income, sales, RFM (recency, frequency, monetary).

Categorical: The categorical data elements are differentiated as follows:

Nominal: These are data elements that can only take on a limited let of values with no meaningful ordering in between. Examples: marital status, profession, purpose of loan.

Ordinal: These are data elements that can only take on a limited set of values with a meaningful ordering in between. Examples: credit rating; age coded as young, middle aged, and old.

Binary: These are data elements that can only take on two values. Example: gender, employment status.

Appropriately distinguishing between these different data elements is of key importance to start the analysis when importing the data into an analytics tool. For example, if marital status were to be incorrectly specified as a continuous data element, then the software would calculate its mean, standard deviation, and so on, this is obviously meaningless.

5. Cloud and big data

Categorization is also known as *coarse classification, classing, grouping, binning* etc can be done for many reasons. For categorical variables, it is needed to reduce the number of categories. With categorization, one would create categories of values such that fewer parameters will have to be estimated and a more robust model is obtained. For continuous variables, categorization may also be very beneficial.

Basic methods to do categorization:

- Equal interval binning
- Equal frequency binning

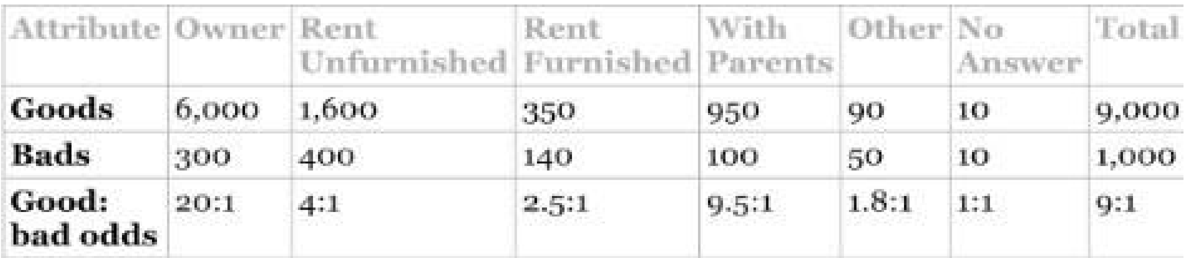
Example: income values 1,000, 1,200, 1,300, 2,000, 1,800 and 1,400.

Equal Interval Binning: it will create two bins with the same range- *Bin1:1,000, 1,500 and Bin2: 1,500, 2,000.*

Equal Frequency Binning: it would create two bins with same number of observations- *Bin1: 1,000, 1,200, 1,3000 and Bin2: 1,400, 1,800, 2,000.*

However, both methods are quite basic and do not take into account a target variable (e.g churn, fraud, credit risk)

Chi-squared analysis is a more sophisticated way to do coarse classification. The table 1.4 shows the example for coarse classifying a residential status variable.



Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1

6. Crowd sourcing

Crowdsourcing: Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models

Kaggle: Kaggle describes itself as "an innovative solution for statistical/analytics outsourcing." That's a very formal way of saying that Kaggle manages competitions among the world's best data scientists. Here's how it works: Corporations, governments, and research laboratories are confronted with complex statistical challenges. They describe the problems to Kaggle and provide data sets. Kaggle converts the problems and the data into contests that are posted on its web site. The contests feature cash prizes ranging in value from $100 to $3 million. Kaggle's clients range in size from tiny start-ups to multinational corporations such as Ford Motor Company and government agencies such as NASA.

As per Anthony Goldbloom, Kaggle's founder and CEO: The idea is that someone comes to us with a problem, we put it up on our website, and then people from all over the world can compete to see who can produce the best solution."

Kaggle's approach is that it is truly a win-win scenario—contestants get access to real-world data (that has been carefully "anonymized" to eliminate privacy concerns) and prize sponsors reap the benefits of the contestants' creativity.

Crowdsourcing is a disruptive business model whose roots are in technology but is extending beyond technology to other areas.

There are various types of crowd sourcing, such as crowd voting, crowd purchasing, wisdom of crowds, crowd funding, and contests.

Take for example:

- 99designs.com/, which does crowdsourcing of graphic design
- agentanything.com/, which posts "missions" where agents vie for to run errands
- 33needs.com/, which allows people to contribute to charitable programs that make a social impact

7. a. Data Discovery

Data discovery allows business users to interact with enterprise data visually to uncover hidden patterns and trends. Data discovery, in the context of IT, is the process of extracting actionable patterns from data. The extraction is general performed by humans or, in certain cases, by artificial intelligence systems.

Lot of buzz in the industry about data discovery, used to describe the new wave of business intelligence that enables users to explore data, make discoveries and uncover insights in a dynamic and intuitive way.

There are two software companies that stand out in the crowd by growing their business are: **Tableau Software and QlikTech International**. They grew through a model called as "land and expand".

tableau

It basically works by getting intuitive software in the hands of some business users to get in the door and grow upward. In order to succeed at the BI game of the "land and expand model" we need a product that is easy to use with lots of effective outputs.

The company's cofounder and chief scientist Pat Harahan: he invented the technology that helped to change the world of animation film. Even though Pat Harahan was not a software BI, he was able to bring the new creative lens in to BI software market. When we have a product that is "easy to use", it is also called as "self-service approach" named by Harahan and his colleagues. Analytics and reporting are produced by the people using the results. IT provides the infrastructure, but business people create their own reports and dashboards.

The important characteristic of rapid-fire BI is the business users, not specialized developers, drive the application. The results that everyone wins. The IT team can stop the backlog of change requests and instead spend time on strategic IT issues. Users can serve themselves data and reports when needed.

Anthony Deighton says that "business intelligence needs to work the way people's mind work. Users need to navigate and interact with data any way they want to- asking and answering questions on their own and in big group or teams".One capability that we have all become accustomed to is search, what many people refers to as "Googling"

Qlicktech has designed a way for users type relevant words or phrases in any order and get instant, associative results.With global search bar the user can search across the entire data set.With search boxes on individual list boxes, users can confine the search to just that field.

b. Open source technology

Definition: Open source software is computer software that is available in source code form under an open-source license that permits users to study, change and improve and at a times to distribute the software.

Origin: The open-source name came out of a 1998 meeting in Palo Alto in reaction to Netscape's announcement of a source code release for navigation (as Mozilla).Although the source code is released, there are still governing bodies and agreements in place. The most prominent and popular example is the GNU General Public License (GPL), which "allows free distribution under the condition that further developments and applications are put under the same license."

As per David Smith - *vice president of marketing at Revolution Analytics in Palo Alto*

In the past, the pace of software development was moderated by a relatively small set of proprietary software vendors. But there are clear signs that the old software development model is crumbling, and that a new model is replacing it.

- The old model's end state was a monolithic stack of proprietary tools and systems that could not be swapped out, modified, or upgraded without the original vendor's support. This model was largely unchallenged for decades.

The status quo rested on several assumptions, including:

1. The amounts of data generated would be manageable
2. Programming resources would remain scarce
3. Faster data processing would require bigger, more expensive hardware

- The sudden increase in demand for software capable of handling significantly larger data sets, coupled with the existence of a worldwide community of open-source programmers, has upended the status quo.
- The old model was top-down, slow, inflexible and expensive. The new software development model is bottom-up, fast, flexible, and considerably less costly.
- A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors. It reflects the old command and control mentality of the traditional corporate world and the old economic order.
- An open-source stack is defined by its community of users and contributors. No one "controls" an open-source stack, and no one can predict exactly how it will evolve. The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

As per Tasso Argyros: *copresident of Teradata*

This is a significant step forward from what was state-of-the-art until yesterday. This means that [in the past] getting data from Hadoop to a database required a Hadoop expert in the middle to do the data cleansing and the data type translation. If the data was not 100% clean (which is the case in most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn't directly access and analyze data in Hadoop clusters. SQL-H, an industry-first, solves all those problems.

8. V's of big data

(i) Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

(ii) Variety – Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(iii) Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

(iv) Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

9. z-scores

1257,1218,1158,5130,3630,21467,1274,4582,6374,2251,2623

Number	Z -score	Outlier?
$1,257	-0.60389	No
$1,218	-0.61086	No
$1,158	-0.6216	No
$5,130	0.088883	No
$3,630	-0.17942	No
$21,467	3.011112	Yes
$1,274	-0.60085	No
$4,582	-0.00914	No
$6,374	0.3114	No
$2,251	-0.42609	No
$2,623	-0.35955	No

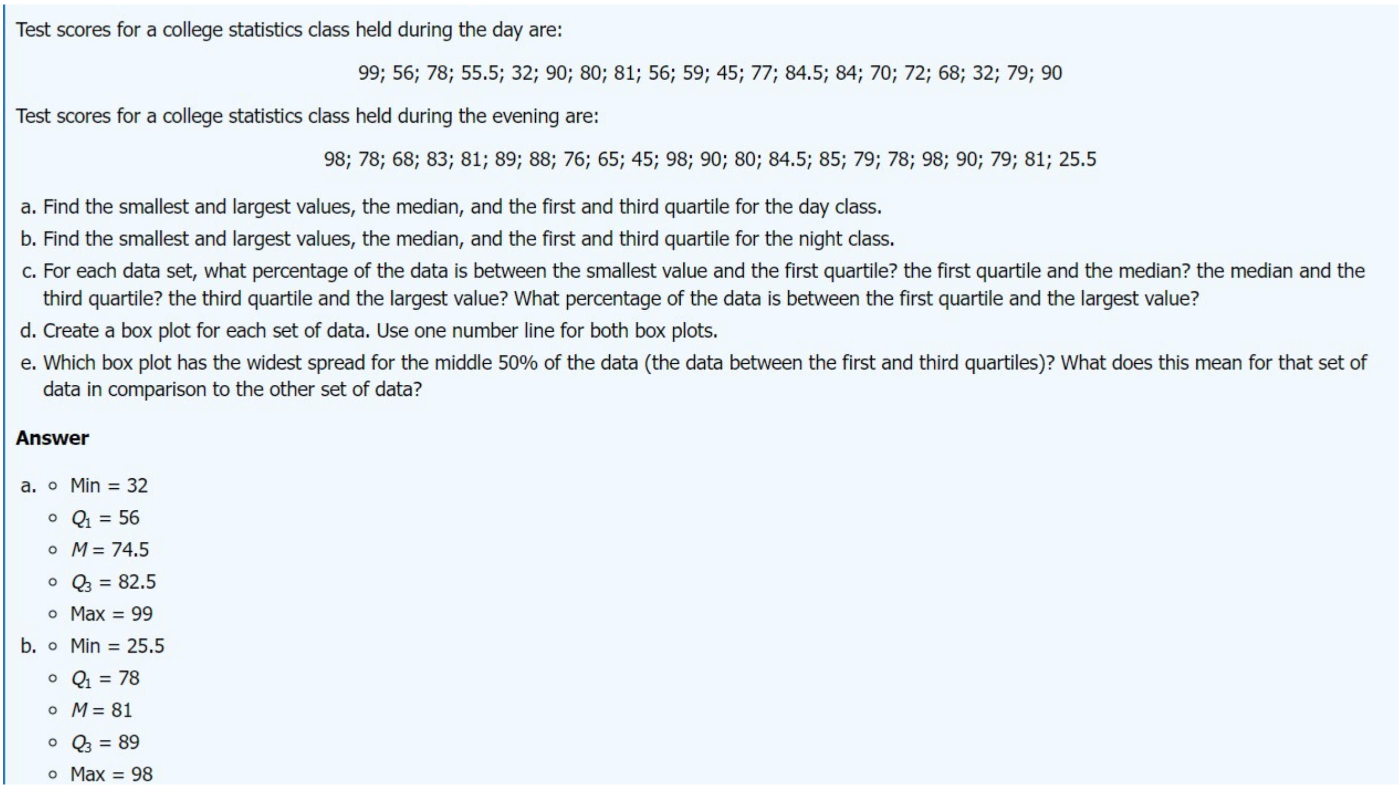
Average= 4633.091

Standard deviation = 5590.596

Z Score=

$$z_i = \frac{x_i - \mu}{\sigma}$$

10. 99;56;78;55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90



Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.

b. Find the smallest and largest values, the median, and the first and third quartile for the night class.

c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?

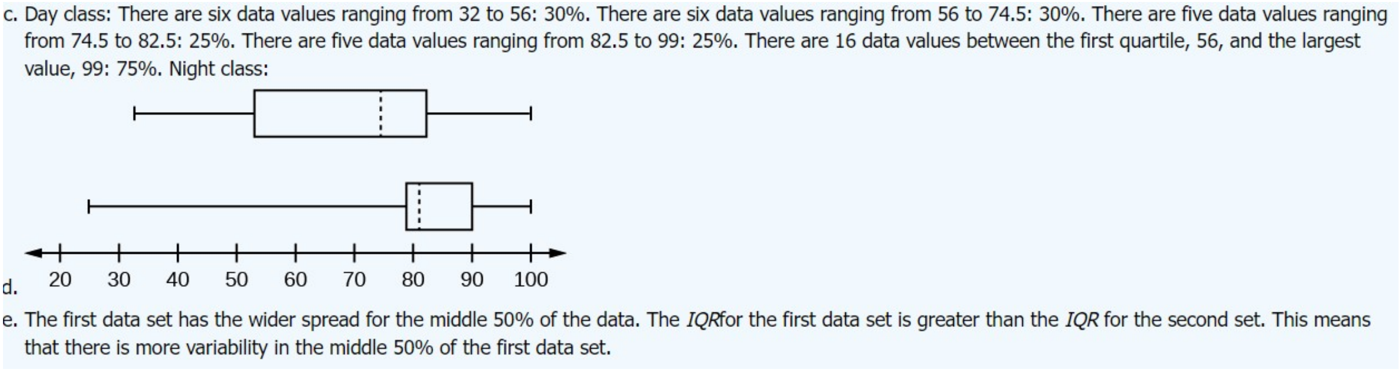
d. Create a box plot for each set of data. Use one number line for both box plots.

e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5
- Q_3 = 82.5
- Max = 99

b.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98



c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

d.

e. The first data set has the wider spread for the middle 50% of the data. The *IQR*for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.