



University of Wales Trinity Saint David – London

Programme: BA Business Management Year 2

Module: Data Handling & Business Intelligence

Module code: LCBB5000

Module Tutor: Anupam Mazumdar

Assignment Title: Coursework (Written Report)

Student Name: Miroslav Gruychev

Student ID: 2021938

Table of Contents

Part 1_Determination of Crime Rate	3
Part 2_Data analysis report	10
SPSS Report.....	10
Evaluation 1_Number of students who like vanilla flavour Ice cream.....	10
Evaluation 2_Number of Male and Female students	11
Evaluation 3_Mean and Median of participants who likes chocolate and strawberry flavour .	12
Part a) Cluster analysis of Ice_cream_csv with SPSS	13
Part b) The most commonly used text mining techniques in business	14
Part c) The advantages and disadvantages of using SPSS and Excel	17
References.....	20

Part 1 Determination of Crime Rate

In business, data analysis refers to the process of screening, processing, and analysing raw data to produce useable, relevant information that can be used to help organisations make informed choices. By providing important insight and data in the form of charts, photos, tables, and graphs, this strategy lessens the risks connected with decision-making. Additionally, the tools and procedures used to analyse data and disseminate conclusions, such as the usage of visualisations and other methods of data exploration are all part of data analytics, which is broader than just data analysis.

For proper procedure of the crime rate determination, it is divided into three segments and they are data pre-processing, analysing and visualizing.

Data pre-processing: There are several procedures used to clean up the extracted raw data before it can be analysed statistically, collectively known as data pre-processing. In order to do effective statistical analysis, pre-processing is necessary. Data cleaning, integration, transformation, and minimization are all part of the process. Sequential pre-processing may occur, and this set of operations may be continued until adequate statistical analysis can be performed. Pre-processing must be done with extreme care to avoid altering the dataset in any manner that might skew the findings of future statistical analyses.

Among the raw data, the relevant data column is Crime rate and crime rate10. For the determination of crime rate, these two columns will be utilized and as a result, the given data is firstly summed and then the average is calculated which would be the final result of the crime rate. And the other variables are reduced for the betterment of calculation.

	CrimeRate	Number of Offences (in million)	CrimeRate10	Number of Offences10 (in million)
	45.5	455	26.5	265
	52.3	523	35.9	359
	56.6	566	37.1	371
	60.3	603	42.7	427
	64.2	642	46.7	467
	67.6	676	47.9	479
	70.5	705	50.6	506
	73.2	732	55.9	559
	75	750	61.8	618
	78.1	781	65.4	654
	79.8	798	71.4	714
	82.3	823	75.4	754
	83.1	831	77.3	773
	84.9	849	78.6	786
	85.6	856	80.6	806
	88	880	82.2	822
	92.3	923	87.5	875
	94.3	943	92.9	929
	95.3	953	94.1	941
	96.8	968	96.2	962
	97.4	974	97.8	978
	98.7	987	99.9	999
	99.9	999	101.4	1014
	103	1030	103.5	1035
	104.3	1043	104.5	1045
	105.9	1059	106.4	1064
	106.6	1066	107.8	1078
	107.2	1072	110.1	1101
	108.3	1083	110.5	1105
	109.4	1094	113.5	1135
	112.1	1121	116.3	1163
	114.3	1143	119.7	1197
	115.1	1151	124.5	1245
	117.2	1172	127.8	1278
	119.7	1197	129.8	1298
	121.6	1216	130.7	1307
	123.4	1234	132.5	1325
	127.2	1272	134.6	1346
	132.4	1324	137.5	1375
	135.5	1355	140.5	1405
	137.8	1378	145.7	1457
	140.8	1408	150.6	1506
	145.4	1454	157.3	1573
	149.3	1493	162.7	1627
	154.3	1543	169.6	1696
	157.7	1577	177.2	1772
	161.8	1618	178.2	1782
Sum	4832	48320	4797.3	47973

In another part the crime rate of youth, male, youth (unemployed) and male (unemployed) are separated for evaluation of their crime rate. These has also been performed under data pre-processing.

CrimeRate (In 100,000)	Crime Rate (in million)	Youth	Youth crime rate	Males	Male crime rate	YouthUnemployment	Unemployed criem rate	MatureUnempl oyment	Unemployed criem rate
45.5	455	135	2.97	965	21.21	80	1.76	22	0.48
52.3	523	140	2.68	1045	19.98	135	2.58	40	0.76
56.6	566	157	2.77	962	17.00	97	1.71	34	0.60
60.3	603	139	2.31	968	16.05	135	2.24	53	0.88
64.2	642	126	1.96	989	15.40	78	1.21	25	0.39
67.6	676	128	1.89	972	14.38	77	1.14	25	0.37
70.5	705	130	1.84	984	13.96	70	0.99	21	0.30
73.2	732	143	1.95	977	13.35	114	1.56	35	0.48
75	750	141	1.88	968	12.91	107	1.43	37	0.49
78.1	781	133	1.70	1024	13.11	99	1.27	27	0.35
79.8	798	142	1.78	969	12.14	94	1.18	33	0.41
82.3	823	123	1.49	948	11.52	124	1.51	50	0.61
83.1	831	135	1.62	986	11.87	77	0.93	27	0.32
84.9	849	121	1.43	964	11.35	84	0.99	29	0.34
85.6	856	166	1.94	973	11.37	72	0.84	26	0.30
88	880	140	1.59	1029	11.69	100	1.14	24	0.27
92.3	923	126	1.37	984	10.66	102	1.11	33	0.36
94.3	943	130	1.38	934	9.90	78	0.83	34	0.36
95.3	953	125	1.31	964	10.12	105	1.10	43	0.45
96.8	968	151	1.56	950	9.81	108	1.12	41	0.42
97.4	974	152	1.56	986	10.12	92	0.94	43	0.44
98.7	987	162	1.64	996	10.09	73	0.74	27	0.27
99.9	999	149	1.49	953	9.54	86	0.86	35	0.35
103	1030	177	1.72	974	9.46	76	0.74	28	0.27
104.3	1043	134	1.28	972	9.32	83	0.80	31	0.30
105.9	1059	130	1.23	1049	9.91	113	1.07	40	0.38
106.6	1066	157	1.47	955	8.96	81	0.76	28	0.26
107.2	1072	148	1.38	998	9.31	84	0.78	20	0.19
108.3	1083	126	1.16	990	9.14	102	0.94	35	0.32
109.4	1094	135	1.23	978	8.94	89	0.81	34	0.31
112.1	1121	142	1.27	956	8.53	116	1.03	47	0.42
114.3	1143	127	1.11	982	8.59	97	0.85	38	0.33
115.1	1151	131	1.14	1038	9.02	142	1.23	42	0.36
117.2	1172	136	1.16	1012	8.63	111	0.95	37	0.32
119.7	1197	119	0.99	938	7.84	92	0.77	36	0.30
121.6	1216	147	1.21	972	7.99	76	0.63	24	0.20
123.4	1234	145	1.18	981	7.95	88	0.71	31	0.25
127.2	1272	132	1.04	953	7.49	83	0.65	32	0.25
132.4	1324	152	1.15	1018	7.69	103	0.78	28	0.21
135.5	1355	125	0.92	985	7.27	130	0.96	58	0.43
137.8	1378	141	1.02	985	7.15	91	0.66	20	0.15
140.8	1408	150	1.07	964	6.85	87	0.62	38	0.27
145.4	1454	131	0.90	969	6.66	79	0.54	35	0.24
149.3	1493	143	0.96	1012	6.78	96	0.64	36	0.24
154.3	1543	124	0.80	966	6.26	77	0.50	35	0.23
157.7	1577	136	0.86	994	6.30	102	0.65	39	0.25
161.8	1618	131	0.81	1071	6.62	102	0.63	41	0.25
			1.47		10.43		1.02		0.36

Data analysis: Data analysis is the mix of data manipulation to make sense of the data and visual representation to demonstrate those evaluations in a much simpler form that can be understood at a look. This definition encompasses a wide range of techniques for obtaining information that may be used in making decisions. Every business need data analysis, whether it is starting up, making marketing choices, carrying on with a given course of action, closing down, or introducing a new product or service. In making the most important decisions, the statistical probabilities and conclusions obtained from data analysis help eliminate any human bias (Northern Illinois

University. 2019). The pre-processed data has been used through different statistical criterial and the final result has been provided as below:

Crime Rate	Crime Rate10	Final Crime Rate
4832	4797.3	102.439

So, the crime rate is found as 102.44

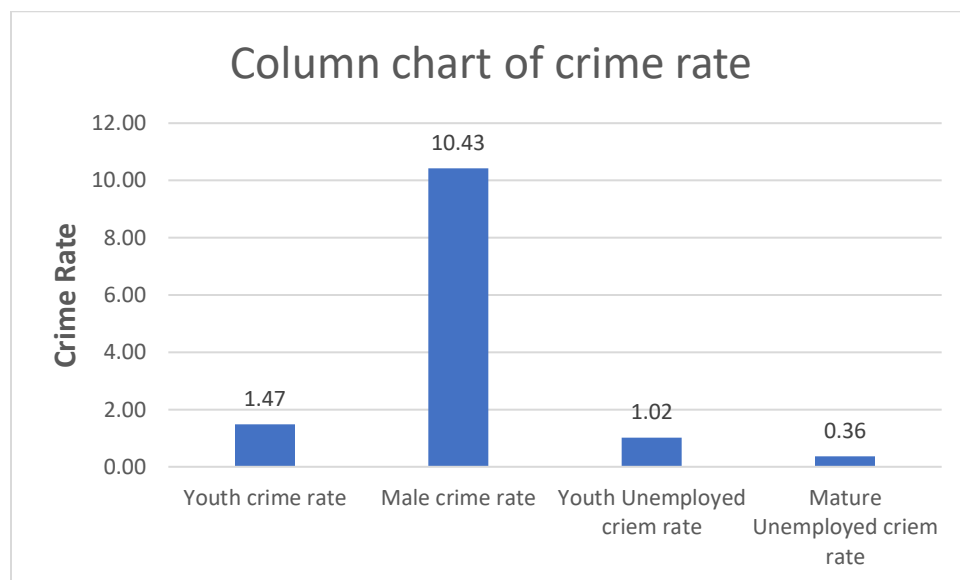
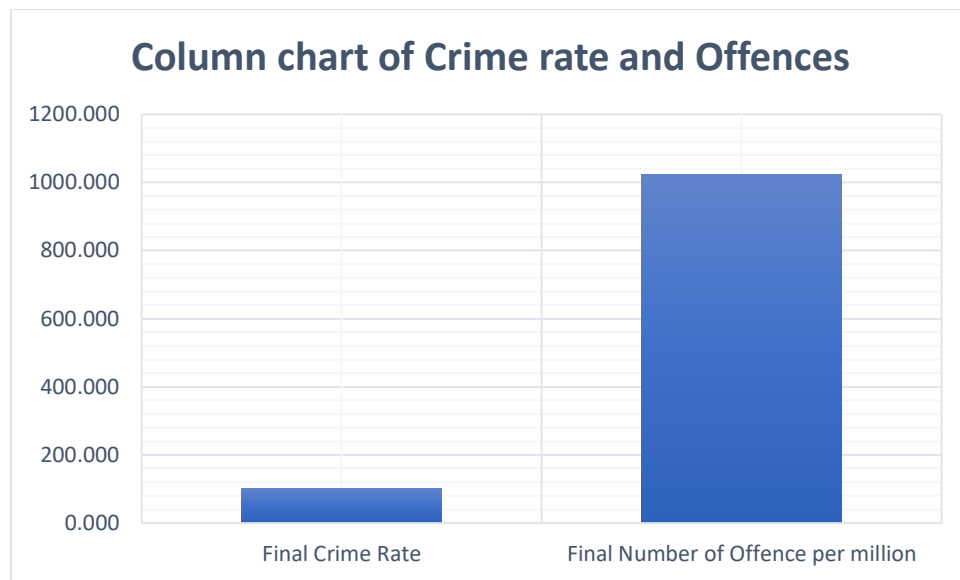
For the number of offences in million, the crime rate and crime rate10 is then processed to evaluated in million. As the crime rate is usually defined in 100 thousand and for converting it into million the crime data is multiplied it with 10. Then the same procedure of crime rate is used for evaluating the number of offences.

Number of Offence	Number of Offence10	Final Number of Offence per million
48320	47973	1024.394

Another variables crime rate

<i>Variables</i>	<i>Value</i>
Youth crime rate	1.47
Male crime rate	10.43
Youth Unemployed crime rate	1.02
Mature Unemployed crime rate	0.36

Data visualization: It's an interdisciplinary field that deals with the graphic depiction of data. The data visualization is presented below with the founded result from the data analysis.



Alternate data analysis method- Pivot Table

Pivot Table analysis

Average of CrimeRate	Average of CrimeRate10	Overall crime rate
102.8085106	102.0702128	102.4393617

Alternate data analysis method- IF Function

Analysis with IF function

Total crime rate	9629.3
Overall crime rate	102.4393617

Current trends in data warehousing, business intelligence and data mining

Data warehousing is the way of transforming data into real-time information for decision-making. It encompasses approaches and tools for storing data in an electronic repository. Data that is time variable, volatile, and subject focused is gathered from numerous sources and turned into homogenous data that can be accessed for analysis and reporting. The primary need in this industry is for high information mining with inexpensive equipment and operational expenses.

Current trends in data warehousing

- **Optimization and Performance:** Comprehensive capability for hardware control of input/output, disk storage, and CPU/memory balance is now incorporated virtually as a matter of capable data warehouse platforms. Several new competitors are concentrating on optimization as a differentiation, and practically every data warehouse provider is now tackling the issue of reducing warehouse space via reduction and usage-based data placement methodologies.
- **Data Warehouse Appliances:** As there are several ways in which firms consider purchasing an appliance, the primary reason is convenience. The supplier creates and validates the setup, harmonizing equipment, software, and applications for consistent performance. The equipment is provided fully assembled and installs quickly.
- **A centralized data warehouse:** Several firms now have several data warehouses, which are typically built on various designs. These are costly and may necessitate the addition of additional staffing resources to achieve evidence of success. To achieve even a moderate level of effectiveness, this necessitates the use of expensive specialist databases and operational personnel. Having all data available in a single service will be one of the main themes this year. This may be accomplished by combining all data into a single repository as well as implementing a large-scale server-less data warehouse. Organizations can also deploy new

data virtualization technique that provide a single data service perspective of many data warehouses, whether on-premise, in the cloud, or a mixture of the two (Foley, n.d.).

- **Mixed Workloads in a Data Warehouse:** Warehouses that supply all workloads must be evaluated for the predictability of mixed workload performance, as neglecting to prepare for mixed workloads is a risk. It will result in rising administrative expenses over time as volume and extra tasks are added, potentially causing serious sustainability difficulties (Anda, n.d.).

Current trends in Business intelligence

- **Cloud-based BI:** The future of business is in the cloud, and big data is following suit. Almost all BI components, including data sources and models, analytics, data storage, and computing capacity, have already been migrated to the cloud. To ensure lower risk and adaptability, businesses require a cloud-connected strategy.
- **Smartphone based Business Intelligence:** When it comes to business analytics, users no longer need to be confined to their desks. Every employee at the organization has a cell phone, and they may do business when away from the office. In today's corporate climate, users may even conduct some or all of their job away from the workplace.
- **Augmented analytics:** Augmented analytics is a critical component of the future of data; it is a disruptive movement that utilizes machine intelligence and artificial intelligence. It alters how data is processed, insights are generated and shared, and data science and machine learning (ML) models are created and formalized (ITProPortal. 2022).

Current trends in Data mining

- **Mining of Spatial and Geographic Data:** The extraction of information from environmental, astronomical, and geographical data is part of the data mining process. It may expose information such as distance and topography, which is mostly employed in geographic information systems and other navigation applications.
- **Consolidation of Vendors:** It is a supply chain technique in which firms restrict their supplier base to a few trusted partners with the purpose of simplifying the procurement process, enhancing the quality of company operations, and, most significantly, saving costs. Significant product convergence in the data mining market is on the horizon as larger database suppliers buy data mining tools companies to supplement their offerings with new capabilities. The

existing, fractured market and diverse set of competitors in the data mining arena are similar to the neighbouring big data vendor ecosystem, which is undergoing consolidation (Data Mining, 2022).

- Automation: It outlines a method for creating a scalable and adaptable machine learning architecture that may be used to automate a variety of data mining jobs. A technique was designed for each activity that automates each phase of the procedures. Earlier iterations of data mining required manual coding by professionals with extensive knowledge of statistics and programming. Modern procedures are substantially automated, with AI/ML replacing the vast majority of formerly manual processes for generating pattern-discovery algorithms. Data mining systems of today often connect ML and large data warehouses to provide enhanced data management features as well as sophisticated data analysis methodologies.
- Data Mining in Embedded Systems: Data mining capabilities are progressively being integrated into a wide range of corporate software applications, from sales forecasting in CRM SaaS platforms to cyber threat detection in intrusion detection/prevention systems. This form of data mining allows prediction skills for a wide range of sectors and opens up new avenues for creating distinctive value (SearchDataManagement, 2022).

Part 2 Data analysis report

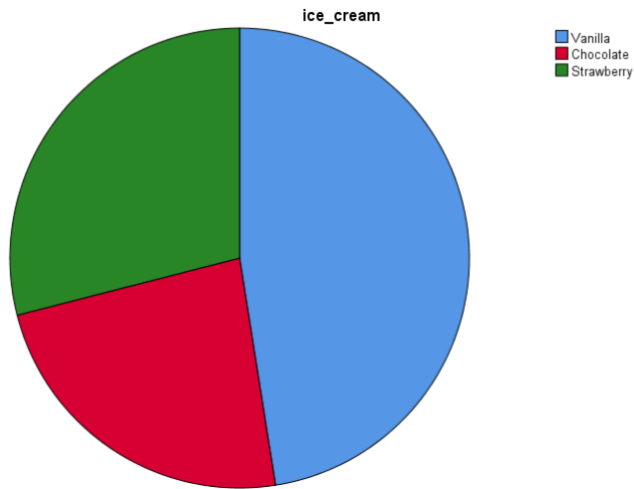
SPSS Report

Evaluation 1 Number of students who like vanilla flavour Ice cream

After analysis in SPSS, the number of students who like different type of flavour can be found.

ice_cream					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Vanilla	95	47.5	47.5	47.5
	Chocolate	47	23.5	23.5	71.0
	Strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

So, the number of students is 95 who like vanilla flavoured Ice cream and the visual demonstration of all flavoured ice cream is presented below:

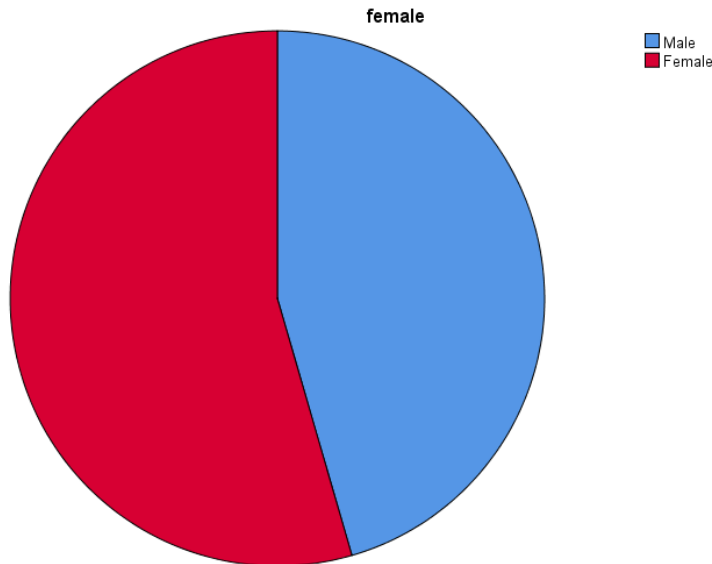


Evaluation 2 Number of Male and Female students

The number of male and female students can be found with the frequency analysis in the SPSS application.

female					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	91	45.5	45.5	45.5
	Female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	

So, the number of male students is 91 and the number of female students is 109 and the graphical presentation is given below:



Evaluation 3 Mean and Median of participants who likes chocolate and strawberry flavour

For calculation of mean and median of students who like chocolate and strawberry, the data first split so that the variables under ice cream can be calculated individually.

ice_cream = Chocolate

Statistics^a

ice_cream		
N	Valid	47
	Missing	0
Mean		2.00
Median		2.00

a. ice_cream =
Chocolate

ice_cream = Strawberry

Statistics^a

ice_cream		
N	Valid	58
	Missing	0
Mean		3.00
Median		3.00

a. ice_cream =
Strawberry

Part a) Cluster analysis of Ice_cream.csv with SPSS

For cluster analysis, firstly the descriptive analysis is performed on ice cream, video and puzzle variables and the below analysis found.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
ice_cream	200	1	3	1.82	.857
video	200	26	74	51.85	9.901
puzzle	200	26	71	52.41	10.736
Valid N (listwise)	200				

The cluster analysis is performed under the K-cluster analysis and the following is the result of that analysis.

Initial Cluster Centers

	Cluster	
	1	2
Zscore(ice_cream)	-.95099	.21587
Zscore(video)	2.23717	-2.30787
Zscore(puzzle)	1.26632	-2.45953

Iteration History^a

Iteration	Change in Cluster Centers	
	1	2
1	2.062	2.152
2	.123	.165
3	.072	.079
4	.074	.083
5	.055	.058
6	.035	.035
7	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 5.992.

Final Cluster Centers

	Cluster	
	1	2
Zscore(ice_cream)	.27364	-.27916
Zscore(video)	.65316	-.66635
Zscore(puzzle)	.71944	-.73397

Number of Cases in each Cluster

Cluster	1	2
	101.000	99.000
Valid	200.000	
Missing	.000	

Part b) The most commonly used text mining techniques in business

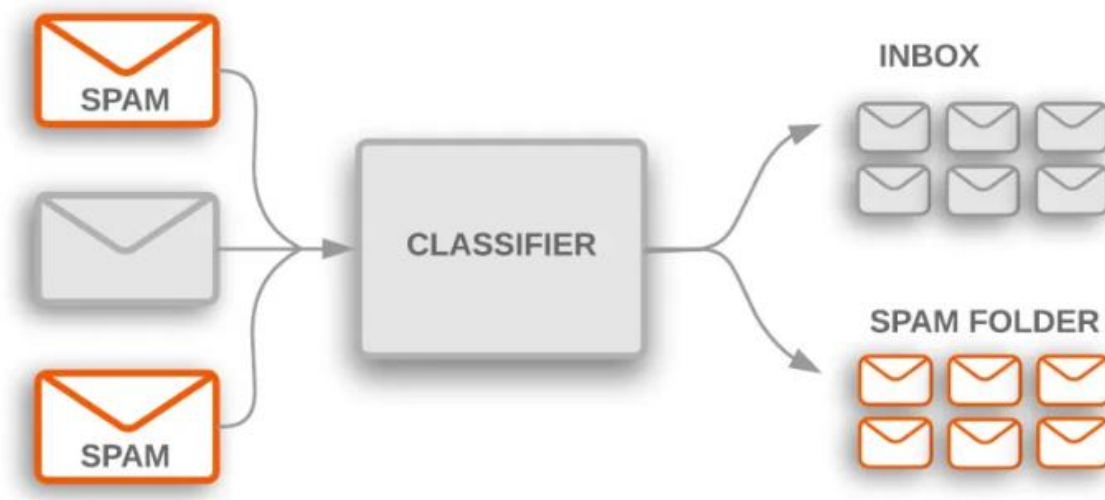
Textual data is encountered by organizations in their day-to-day operations. This data comes from a number of sources and may be available, but it stays untapped owing to a lack of knowledge of

an organization's information richness or a lack of technique or technology to evaluate this data and get relevant insight. Each type of information is a valuable asset that may provide insight into a company's operations by allowing it to make decisions based on that information. Such data might include details on their customers, partners, and rivals. The data might provide insight into their customers and how to improve connections with their partners. This can assist them in staying ahead of their competition. As a result, it is critical to employ approaches that can capitalize on this potential by revealing hidden value in this data. And that is where text mining techniques come in handy, as they may aid in the discovery of important and fascinating knowledge from this data. Such strategies are used by businesses to examine consumer and competition data in order to increase competitiveness. The short description of some most commonly used text mining techniques in business are given below:

Text Classification:

It is a machine learning technique that categorizes open-ended text into one of many predefined categories. Almost any form of data may be organized, arranged, and classified using it. Text is one of the most common types of unstructured data, accounting for more than 80% of all information. Text classifiers may be used by corporations to automatically organize all information and more in a timely and cost-effective manner. As a consequence, businesses may save time researching text data, automate business processes, and make data-driven business decisions.

Gmail spam classifier employs filters based on TensorFlow, the company's proprietary machine learning platform that has trained machine learning algorithms to stop an additional 100 million spam messages every day. It can recognize trends over time depending on what particular Gmail users deem spam (Ganesan, 2020).



Source: Developers.google.com

Text extraction

Text extraction is the extraction of data from textual data such as keywords and employs machine learning to automatically scan text and extract important words and phrases from unstructured information such as news articles, surveys, and customer service requests. Word clouds and insights on brand association are frequently created via keyword extraction. Text extraction may be used by marketing teams to uncover commonly talked themes (Ontotext, n.d.).

Below is an example of text extraction where the relevant data or the necessary data has been collected from a given text and this process is called text extraction.

Results	
TAG	VALUE
brand	iView
display	11.6
cpu	Intel Bay Trail Z3735F
memory	2GB
disk	32GB

iView Maximus II 11.6 Laptop,
Touchscreen, 2-in-1, Windows
10, Intel Bay Trail Z3735F
Processor, 2GB RAM, 32GB
Storage

Source: Ontotext

Text clustering

It is the use of cluster analysis on text-based materials. It understands and categorizes unstructured textual material using machine learning and natural language processing. Clusters are less accurate than classification algorithms, but they are easier to build. It makes it easy for clever algorithms to gather data and generate forecasts without the need for machine learning techniques (Team, 2018).

Customers and target audiences are talking about the goods and features they want all over the web, but it's historically tough to gather all of the data and convert it into a meaningful report. Depending on a few reviews and forum discussions, it's difficult to tell how many people actually desire a feature. However, with text clustering, all of the customer and target audience evaluations can be examined and utilized to construct a roadmap of features and goods that they would enjoy.

Part c) The advantages and disadvantages of using SPSS and Excel

IBM's SPSS (Statistical Package for the Social Sciences) is a tool for performing statistical data analysis. It provides a fast-visual modelling environment for models ranging from the most basic to the most sophisticated.

Advantages

- SPSS is a commonly used program in business research and other social science domains.
- It is interoperable with a wide range of other data management applications. As a result, it enables data flow from one program to another.
- It enables the processing of enormous volumes of data, the use of bigger samples, and the inclusion of additional variables.
- It enables the shift of focus from mechanical computation activities to intellectual tasks such as process judgments, includes detailed, and thoughtful evaluation (SurveySparrow, 2022).

Disadvantages

- It cannot be used to examine large amounts of data.
- The graphics are not suitable for publishing.
- Many strategies lack information on effect magnitude and confidence intervals.
- Algorithm documentation might be difficult or hard to come by at times.
- Pricey in comparison.

Microsoft Excel has long been regarded as one of the most effective systems for spreadsheet production and analysis. With a plethora of built-in capabilities to make sophisticated computations easier, it has become one of the most popular data analysis applications.

Advantages

- Excel can arrange vast volumes of data into neat, logical spreadsheets and graphics.
- Excel crunches numbers almost fast, making bulk computations far more convenient than figuring things out individually.
- To generate more visually dynamic reports, Excel spreadsheets may be readily incorporated to Word pages and PowerPoint presentations.

Disadvantages

- Excel has no way of checking for human mistake when entering data.

- It might take a long time to manually enter data into Excel.
- To use the cloud-based version of Excel, you must have an Office 365 subscription (Techwalla. 2015).

References

Anda, P.K.C., What, Bila. n.d. *Trends In Data Warehousing - ppt download*. [online] slideplayer.com. Available at: <https://slideplayer.com/slide/13154727/> [Accessed 14 Apr. 2022].

Data Mining - Applications & Trends. 2022. Data Mining - Applications & Trends. [ONLINE] Available at: https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm. [Accessed 17 April 2022].

Foley, J. n.d. *Oracle BrandVoice: The Top 10 Trends In Data Warehousing*. [online] Forbes. Available at: <https://www.forbes.com/sites/oracle/2014/03/10/the-top-10-trends-in-data-warehousing/?sh=754480a13737> [Accessed 17 Apr. 2022].

Ganesan, K. 2020. *5 Examples of Text Classification in Practice*. [online] Kavita Ganesan, PhD. Available at: <https://kavita-ganesan.com/5-real-world-examples-of-text-classification/#.YlTEZ2hBzIU> [Accessed 12 Apr. 2022].

ITProPortal. 2022. The future of business intelligence: Six key trends | ITProPortal. [ONLINE] Available at: <https://www.itproportal.com/features/the-future-of-business-intelligence-six-key-trends/>. [Accessed 14 April 2022].

Northern Illinois University. 2019. *Data Analysis*. [online] Hhs.gov. Available at: https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html.

Ontotext. n.d. *What is Information Extraction?* [online] Available at: <https://www.ontotext.com/knowledgehub/fundamentals/information-extraction/> [Accessed 12 Apr. 2022].

SearchDataManagement. 2022. Top Trends in Big Data for 2022 and Beyond. [ONLINE] Available at: <https://www.techtarget.com/searchdatamanagement/feature/Top-trends-in-big-data-for-2021-and-beyond>. [Accessed 14 April 2022].

SurveySparrow. 2022. *What is SPSS? Definition, Features, Types, and Use Cases*. [online] Available at: <https://surveysparrow.com/blog/what-is-spss/#section7> [Accessed 12 Apr. 2022].

Team, E. 2018. *What is Text Clustering?* [online] insideBIGDATA. Available at: <https://insidebigdata.com/2018/07/26/what-is-text-clustering/>.

Techwalla. 2015. *Disadvantages of Using Microsoft Excel* | *Techwalla.com*. [online] Available at: <https://www.techwalla.com/articles/disadvantages-of-using-microsoft-excel>.