Data Driven Decision Making using Data Science and Analytics

Stephen Opiyo

Presentation to Nairobi Women in Machine Learning and Data Science

May 26th, 2018

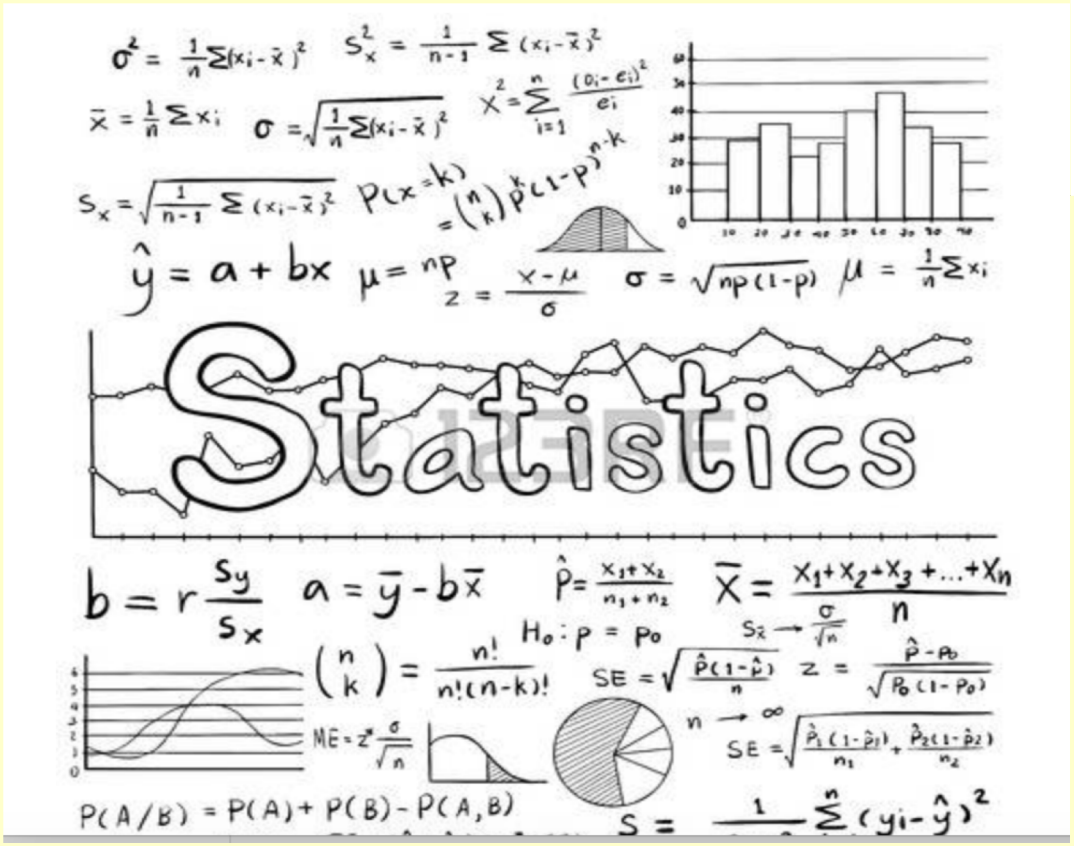
Overview

- Introduction
  - Data Science
  - Analytics
- Data Science and Analytics Projects
- Questions

What is Data Science?

$$\mathrm{Cov}(A,B) = \frac{1}{n-1}\left((a_1-\mu_A)(b_1-\mu_B)+\ldots+(a_n-\mu_A)(b_n-\mu_B)\right).$$

$$\mathrm{Var}(A) = \frac{1}{n-1}\left((a_1-\mu_A)^2+\ldots+(a_n-\mu_A)^2\right).$$



$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0$$

$$0 - 0 - \phi 1 = 0$$

$$\tilde{\boldsymbol{\beta}} = \sum_{k=1}^{m} l_k^{-1}\mathbf{a}_k\mathbf{a}_k'\mathbf{X}'\mathbf{y}$$

$$\begin{aligned}\mathrm{cov}(\alpha_1'\mathbf{x},\alpha_2'\mathbf{x}) &= \alpha_1'\boldsymbol{\Sigma}\alpha_2 = \alpha_2'\boldsymbol{\Sigma}\alpha_1 = \alpha_2'\lambda_1\alpha_1' \\ &= \lambda_1\alpha_2'\alpha = \lambda_1\alpha_1'\alpha_2 = 0\end{aligned}$$

What is Data Science?

- Data - driven science
- Interdisciplinary field of discovering knowledge/insights from data
- Quantitative/qualitative science, information technology, business and communications
- Analytics is the engine of Data Science

What is Analytics?

- Analytics allow us to derive **value** from data by answering **four** key questions:
  - What happened?
  - Why did it happen?
  - What’s is happening now?
  - What might happen in the future?

What is Analytics?

- **Analytics** is a process by which you take raw material **(data)** and convert it into **relevant insights** that can **inform** organization, **improve performance** and **guide strategy.**

Types

- **Descriptive analytics:** Describing what happened (What is the problem). It does not employ statistical algorithms.

- **Diagnostics analytics:** Diagnosing what is wrong

- **Predictive analytics:** Predicting what is ahead by using collected data to train models that will predict what is in the future.

- **Prescriptive analytics:** anticipates what will happen and what should be done to benefit from the predictions.

The raw material of analytics - Data

- In general, data is any set of characters that has been gathered and translated for some purpose, usually analysis. It can be any character, including text and numbers, pictures, sound, or video.

- Data contains variables (key performance indicators)
  - Examples of variables: Age, income, education, etc.

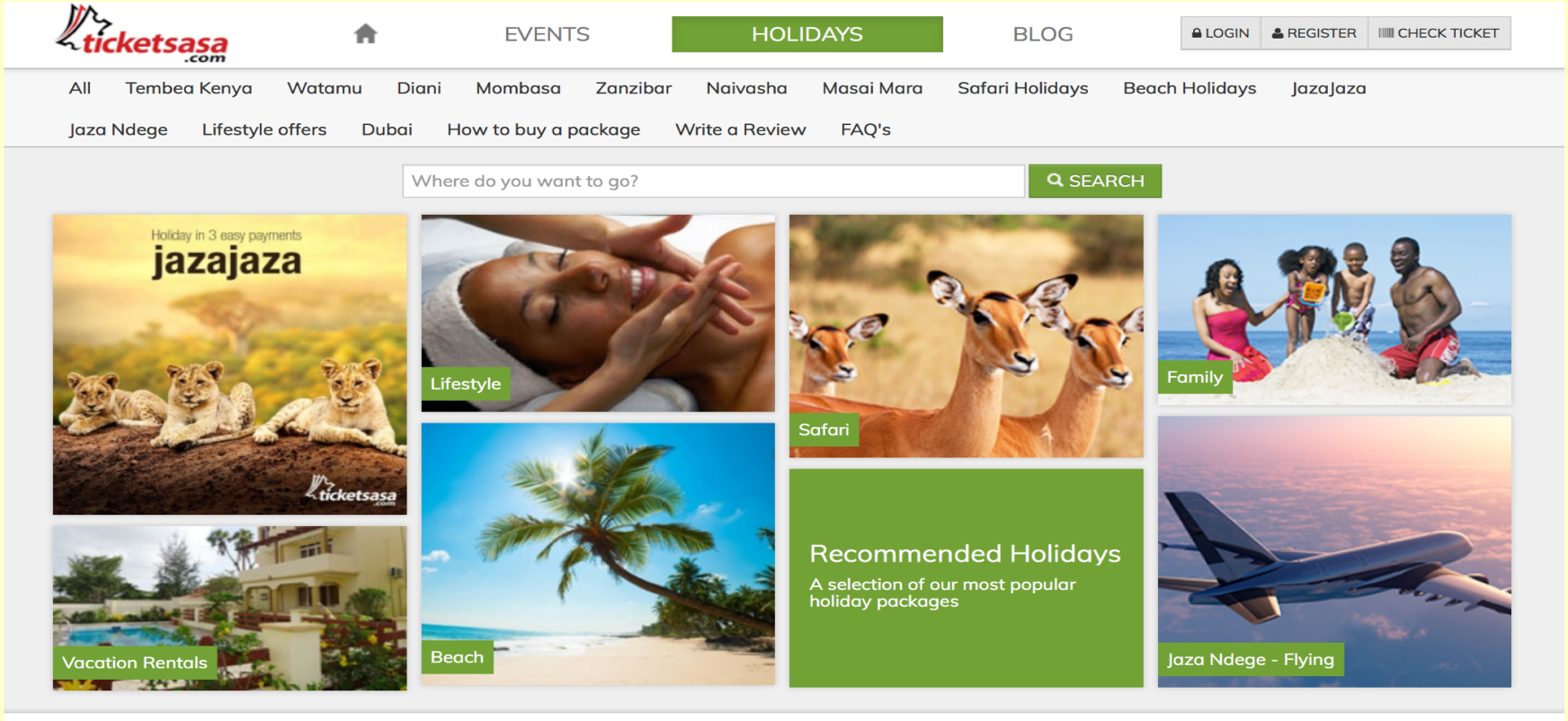
Type of data: Structured

- Structured data is highly organized and located in a fixed field within a defined record or file.

Age	Duration	Years_Matur	Ranking
44	13	15	1
42	14	24	1
40	14	16	1
44	14	17	1
55	13	14	1
48	13	16	1
47	13	16	1
38	13	24	1
45	13	20	1
46	14	16	1
44	13	18	1
55	13	16	1
47	13	16	1
42	13	20	1
41	14	24	1
52	13	24	1
44	12	16	1
34	13	20	1
41	12	17	1
43	8	10	1
52	10	12	1
47	13	15	1
51	12	16	1

Types of data: Unstructured

Unstructured data is everything else that can't be easily inserted into fields, rows or columns. It is often text heavy may contain dates, numbers, different types of data such as images.



Type of data: Semi-structured

Semi-structured data is a hybrid of unstructured and structured data.

ID	text	favorited	favoriteCoun	replyToSN	created	truncated	replyToSID	nuber	replyToUID	screenName	retweetCoun	isRetweet
1	RT @Ma3Ro	FALSE	0	NA	11/10/17 21:25	FALSE	NA	929097728042422000	NA	YongaJnr	26	TRUE
2	RT @Ma3Ro	FALSE	0	NA	11/10/17 21:12	FALSE	NA	929094558817890000	NA	AshMutiso	26	TRUE
3	RT @Ma3Ro	FALSE	0	NA	11/10/17 19:25	FALSE	NA	929067594392403000	NA	isaacyego51	26	TRUE
4	RT @Ma3Ro	FALSE	0	NA	11/10/17 18:59	FALSE	NA	929060925193244000	NA	ItsThatGirlEl	26	TRUE
5	RT @Ma3Ro	FALSE	0	NA	11/10/17 18:58	FALSE	NA	929060789369147000	NA	Kaka_Dubu	26	TRUE
6	RT @suemc_	FALSE	0	NA	11/10/17 18:53	FALSE	NA	929059492578422000	NA	tomwaswa	56	TRUE
7	RT @Kenyan	FALSE	0	NA	11/10/17 18:41	FALSE	NA	929056519806742000	NA	AludaJoyce	5	TRUE
8	Kenya Orient	FALSE	0	NA	11/10/17 18:39	FALSE	NA	929055991341174000	NA	WaweruMw	0	FALSE
9	RT @suemc_	FALSE	0	NA	11/10/17 18:25	FALSE	NA	929052532776857000	NA	FellohMuas	56	TRUE
10	RT @Ma3Ro	FALSE	0	NA	11/10/17 18:01	FALSE	NA	929046402361167000	NA	EldoretBeaut	26	TRUE
11	RT @Ma3Ro	FALSE	0	NA	11/10/17 17:32	FALSE	NA	929039053470818000	NA	AsadRimsy	26	TRUE
12	RT @suemc_	FALSE	0	NA	11/10/17 17:31	FALSE	NA	929038792165556000	NA	mbugga	56	TRUE
13	RT @suemc_	FALSE	0	NA	11/10/17 17:21	FALSE	NA	929036399189417000	NA	AdmaniAma	56	TRUE
14	RT @Ma3Ro	FALSE	0	NA	11/10/17 17:08	FALSE	NA	929033047130935000	NA	DennohCheru	26	TRUE
15	RT @suemc_	FALSE	0	NA	11/10/17 17:05	FALSE	NA	929032311093555000	NA	Pete_Wanyo	56	TRUE
16	RT @Ma3Ro	FALSE	0	NA	11/10/17 16:31	FALSE	NA	929023598982864000	NA	HillaryJrn	26	TRUE
17	@citizentvke	FALSE	0	citizentvkeny	11/10/17 16:25	TRUE	9.2902E+17	929022136685072000	70394965	omolomoses	0	FALSE
18	RT @suemc_	FALSE	0	NA	11/10/17 16:21	FALSE	NA	929021122947047000	NA	uhsam357	56	TRUE
19	Insurance fir	FALSE	0	NA	11/10/17 16:21	FALSE	NA	929021082815709000	NA	OngusoN	0	FALSE
20	RT @Ma3Ro	FALSE	0	NA	11/10/17 16:13	FALSE	NA	929019094472503000	NA	NgetaDiana	26	TRUE
21	RT @suemc_	FALSE	0	NA	11/10/17 16:12	FALSE	NA	929018933268615000	NA	EstaStar	56	TRUE
22	RT @Ma3Ro	FALSE	0	NA	11/10/17 16:01	FALSE	NA	929016201606557000	NA	MPnjenga9	26	TRUE

Source of data

- Internal data source and external data source.

- Internal data source is owned and controlled by the organizations.

- As a rule of thumb, **internal data** is usually easier and cheaper to access because it is controlled by the organizations. They might include claim history, financial records, customer feedback, transaction history, employee , etc.

Source of data

- External data is any data that exists outside the organizations which is held either publicly or privately by another organization.

- External data might include weather data, social media profile data, trend data or government-held data such as census information.

Source and type of data

	Structured	Semi-Structured	Unstructured
Internal	• Point of sale • Financial data • Customer data • HR records	• Photos or graphics with tags or categories • Videos with tags or categories • Email	• Website • Text files • Photos • Audio • Social media
External	• Market research • GPS position data • Sensor data • Weather data	• Tagged photos • Organised graphics • Tagged videos • Categorised text	• Website • Text files • Photos • Audio • Social media

Overview

- Introduction
  - Data Science
  - Analytics
- Data Science and Analytics Projects
- Questions

Insights from Insurance Data From Kenya

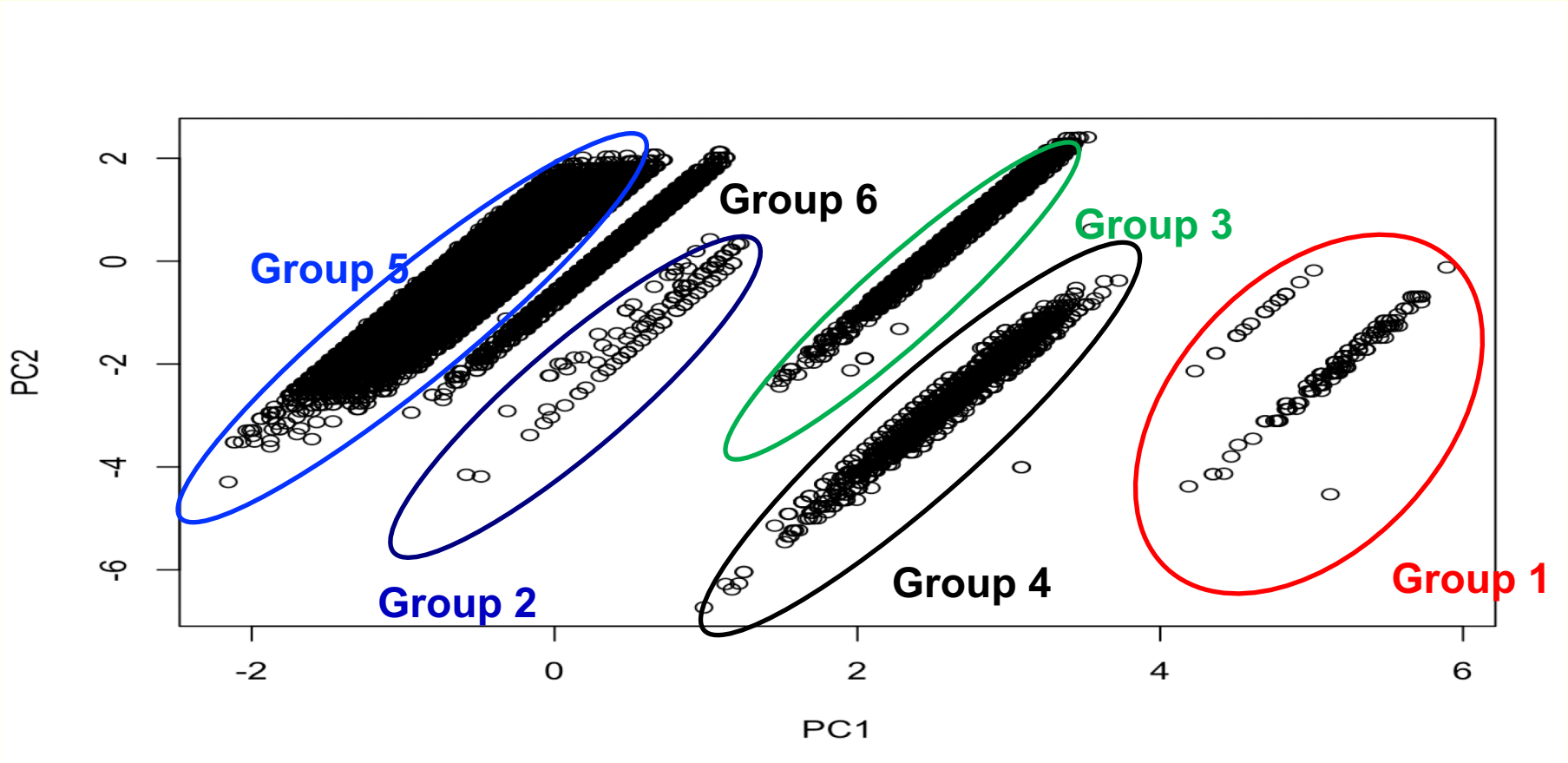
Organization question	Raw material (Data)	Data source	Data type
Can we reduce lapse Rate of lapse is 36%?	Yes	Internal	Semi-structured

- Format: Excel
- Data type: Semi-structured
- Data source: Internal
- Observations: 53,776
- Variables: 22

Sample data

Status	PeriodFrom	PeriodTo	SumAssure	TotalPremiumPayable	PremiumFrequency	ProposalSerial	PayMode
Policy	6/5/13	6/5/26	1244536	10,435.00	MONTHLY	18111	Direct Debit
Policy	6/13/17	6/13/30	1843611	15,648.00	MONTHLY	IL078095 FU	Direct Debit
Policy	12/1/14	12/1/24	3459459	30,075.00	MONTHLY	34361	Check Off
Policy	8/31/15	8/31/28	1732164	14,035.00	MONTHLY	048275 FU	EFT
Policy	7/27/16	7/27/26	247935	602.00	MONTHLY	IL063704FU	Check Off
Policy	7/27/16	7/27/26	247998	3,008.00	MONTHLY	IL063703FU	Check Off
Policy	7/8/16	7/8/26	777930	10,025.00	MONTHLY	IL050693LU	Direct Debit
Policy	5/10/16	5/10/26	418323	5,116.00	MONTHLY	IL056221FU	Direct Debit
Policy	5/10/16	5/10/26	262332	247.00	MONTHLY	IL056218FU	Direct Debit
APL	6/29/13	6/29/26	297805	2,506.00	MONTHLY	10149	M-Pesa
APL	6/29/13	6/29/23	232562	501.00	MONTHLY	7249	M-Pesa
Policy	6/10/13	6/10/28	2968460	20,050.00	MONTHLY	IL017017FU	M-Pesa
Policy	7/11/14	7/11/29	864147	6,159.00	MONTHLY	IL015245FU	Direct Debit
Policy	11/10/16	11/10/31	2170271	15,040.00	MONTHLY	IL027954LU	Cheque
APL	5/2/13	5/2/23	1063123	10,025.00	MONTHLY	IL 016507 FU	Direct Debit
Policy	8/30/16	8/30/31	3851429	25,000.00	MONTHLY	IL064345FU	EFT
Policy	7/16/15	7/16/25	1676973	20,050.00	MONTHLY	IL 041931 FU	EFT
Policy	3/5/15	3/5/24	398514	5,144.00	MONTHLY	IL 040903 FU	M-Pesa
Policy	3/5/15	3/5/25	268823	3,096.00	MONTHLY	IL 040837 FU	M-Pesa
Lapsed	3/5/15	3/5/25	536553	5,230.00	MONTHLY	IL 040905 FU	M-Pesa
Lapsed	12/20/16	12/20/26	746182	7,200.00	MONTHLY	IL056041LU	Direct Debit
Policy	2/15/17	2/15/27	211559	2,069.00	MONTHLY	IL067925LU	EFT
Policy	1/9/17	1/9/27	278384	3,008.00	MONTHLY	IL070330LU	EFT
Policy	5/28/17	5/28/27	1006637	10,000.00	MONTHLY	IL060621LU	Direct Debit
Policy	8/28/15	8/28/25	332381	3,200.00	MONTHLY	021917 LU	Check Off
Policy	8/28/15	8/28/28	390171	3,200.00	MONTHLY	021924 LU	Check Off
Policy	11/17/17	11/17/27	74626	201.00	MONTHLY	IL079977FU	M-Pesa

PCA of agents



Average of variables for each group

Groups

Group 1: Agent = 1 policy (540)

Group 2: Agent = 2 (287)

Group 3: Agent = 3 -19 (845)

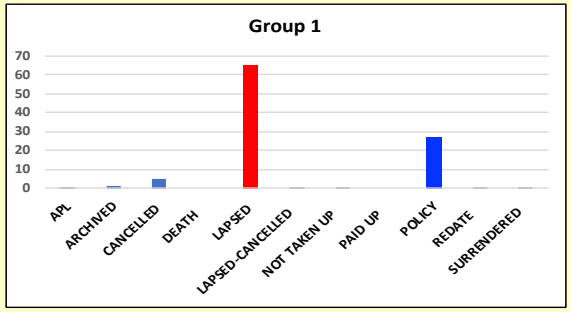
Group 4: Agent = 20 – 100 (329)

Group 5: Agent = 101 -296 (329)

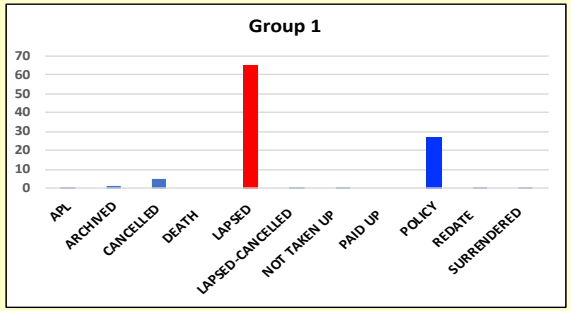
Group 6: Agent = 300 -680 (14)

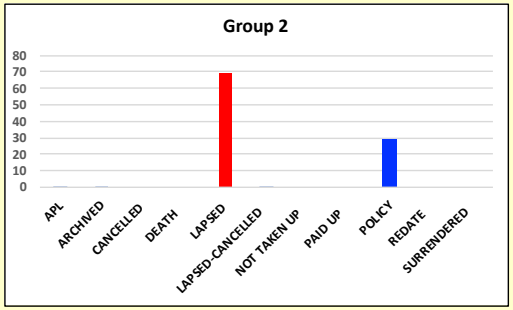
Group	Age	Term	Sum assured	Total Premium
Group 1	30.0	10.40	355172	4229
Group 2	32.0	10.52	361022	4016
Group 3	34.3	10.72	359355	4100
Group 4	35.6	11.26	497963	4867
Group 5	36.1	11.72	565714	5613
Group 6	37.8	11.75	799034	9415

Group 1

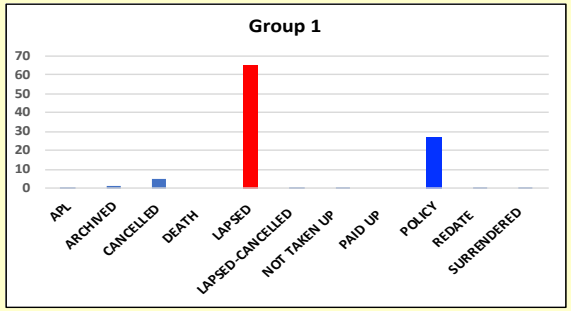

Group 1
70
60
50
40
30
20
10
0
APL
ARCHIVED
CANCELLED
DEATH
LAPSED
LAPSED-CANCELLED
NOT TAKEN UP
PAID UP
POLICY
REDATE
SURRENDERED

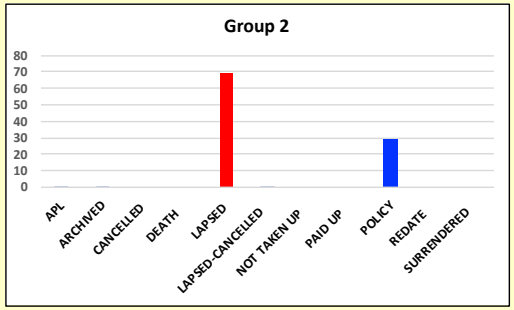
Group 2 added

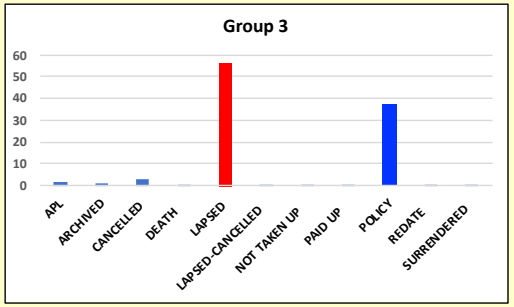




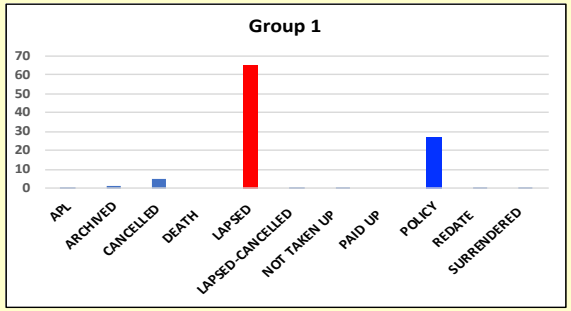
Group 3 added

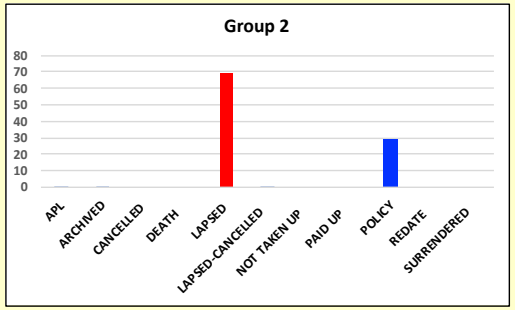


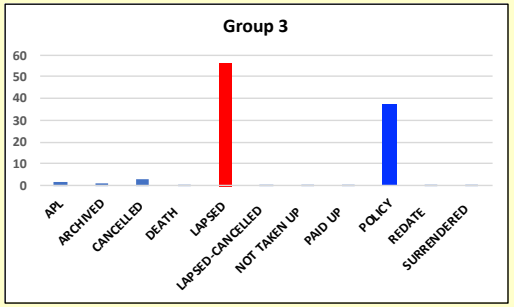


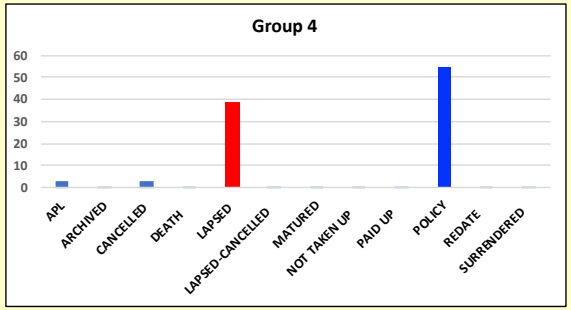


Group 4 added

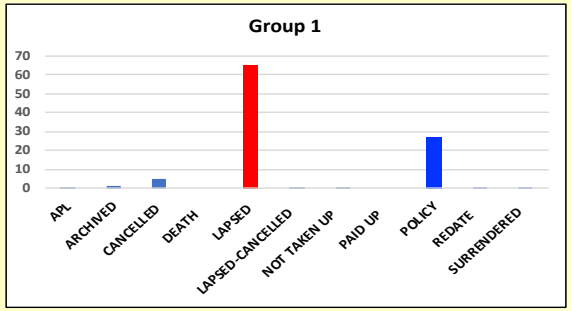


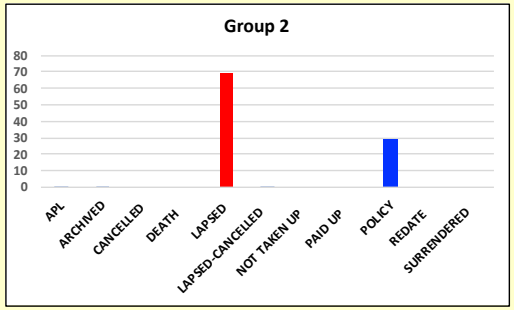


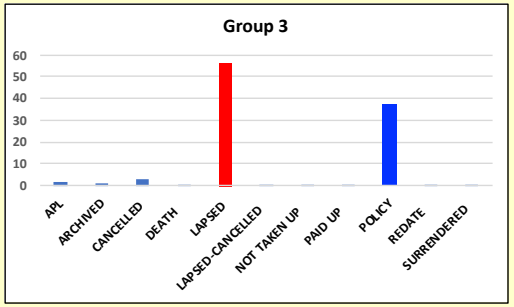


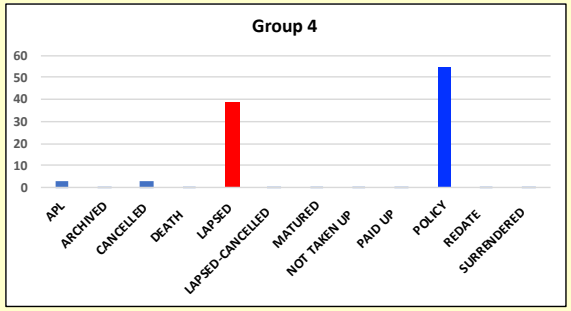


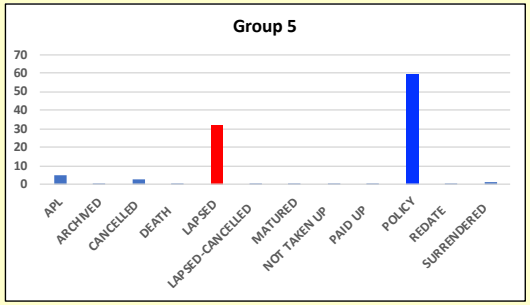
Group 5 added



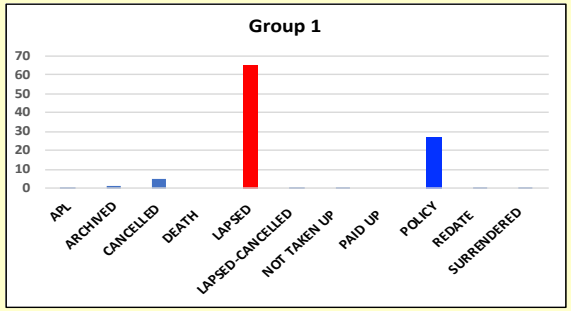


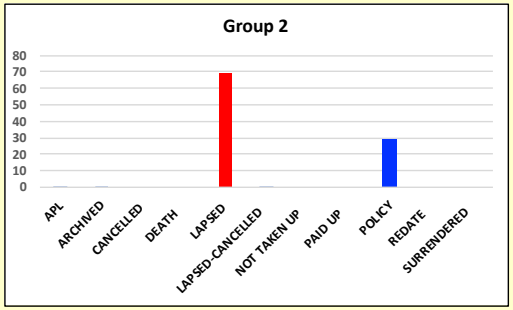


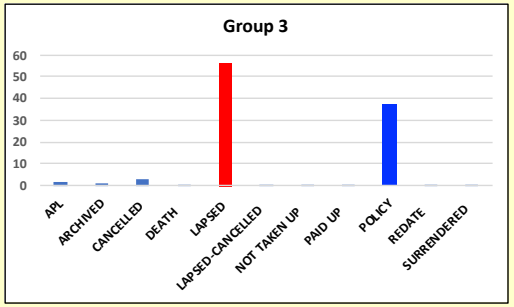


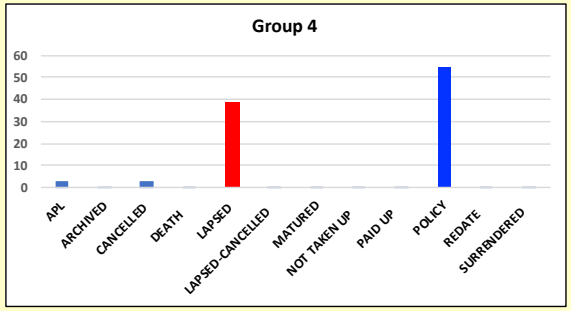


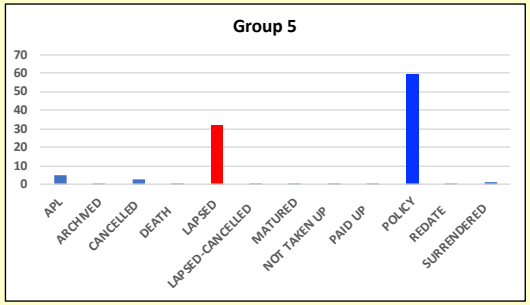
Group 6 added

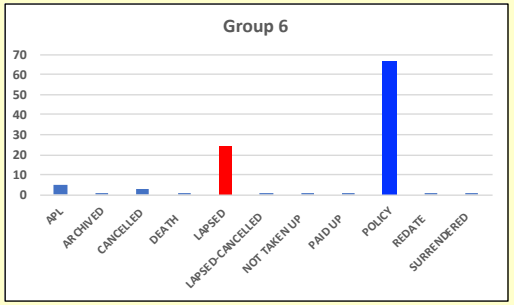












Recommendations

- Fire all or 90% of agents in Groups 1 and 2 (1,367)
- For group 3 work with only agents with less than 10% lapse rate
- For groups 4, 5, and 6 work with only agents with less than 15% lapse rate
- Predict people are most likely to lapse
- Predict bad agents before giving them contracts

Real Estate Analytics

Organization question	Raw material (Data)	Data source	Data type
Can we predict people who are most likely to sell homes within 12 months?	Yes	External 3rd party	Unstructured

- Data type: Unstructured
- Data source: External
- Observations: 130,000,000 houses (USA)
- Variables: 2048

3rd Party sample data

|CensusFIPS	PlaceCode||	CensusTract||	CensusBlockGroup||	CensusBlock||	ParcelNumberRaw||	ParcelNumberFormatted||	ParcelNumberYearAdded||	ParcelNumberAlternate||	ParcelMapBook||Par
METROPOLI	DIVISION||V	PALM	BEACH||936	N	FLAGLER	DR	APT	102||917||I	PALM
BEACH||FL|	CLUB	TOWERS	CONDO	UNIT					
SEC	2||||||||	GLEN	WEST	SEC	02|||0||6	HELEN||HEL	MELLOWGR(	LN||18615|||	MELLOWGROVE||LN|||||||
SPRING||TX||77379||4091||C008|||0||1									
METROPOLI	DIVISION||E	BEACH-DELR	BEACH||902	CORAL	LAKES	DR||12759|	LAKES||DR|	BEACH||FL|	LAKES
IV	LT	130||42E||4	LAKES	04|||0|||					
METROPOLI	DIVISION||V	Palm	Beach-Boca	Raton-Delray	Beach	FL||92990||	SUMMERTIN	LN||10615|	PALM
BEACH||FL|	ESTS	PL	NO	9					
JOSE	&	MARIA||JUA	MARIA||MA	NEW	JERSEY	ST||2017||	JERSEY||ST|||||||	BAYTOWN||TX||77520||6047||C016||||||1||||||||||||||||||||||||||||||||								
METROPOLI	DIVISION||E	BEACH-DELR	BEACH||902	CORAL	LAKES	DR||12771|	LAKES||DR|	BEACH||FL|	LAKES
IV	LT	133||42E||4	LAKES	04|||0|||					
METROPOLI	DIVISION||V	COMMUNITY	67TH	ST	N||12357||	PALM	BEACH||FL|	N	241.5
FT	OF	S	4103FT	OF					
FT	OF	LTS	4	5	&	6	BLK	18	BURRIS
SEC	1|||||||||	SEC	01|||0||18	BRIAN||BRI					
METROPOLI	DIVISION||E	BEACH-DELR	BEACH||902	CALLAWAY	GARDENS	RD||12257|	GARDENS||	BEACH||FL|	GREENS
AT	PIPERS	GLEN	LOT	82B||42E||					
SYLVIA	E||WILLIAM	SYLVIA	E||SYLVIA||	ANGELES||C	OLIVE	GLEN	LN||11911|	GLEN||LN||	ANGELES||CA||90047||5359||C006||||||||||||||||||||||||||||||||||||||||||								
OUTLOT	23||||||||	OUTLOT	23|||0||5	SANDRA	BUI||SANDF	BEND||4815	BOX	20102|||||	BOX
20102|||||	LAND||TX||77496||0102||B035||||||||||||||								
1|||||||||	VALLEY	SEC	01|||0||1	SARA||SARA	SANDYGATE	LN||7606||||	SANDYGATE||LN||||||||	HOUSTON||TX||77095||4018||R074|||0|||||||||||||||||||

http://offrs.com

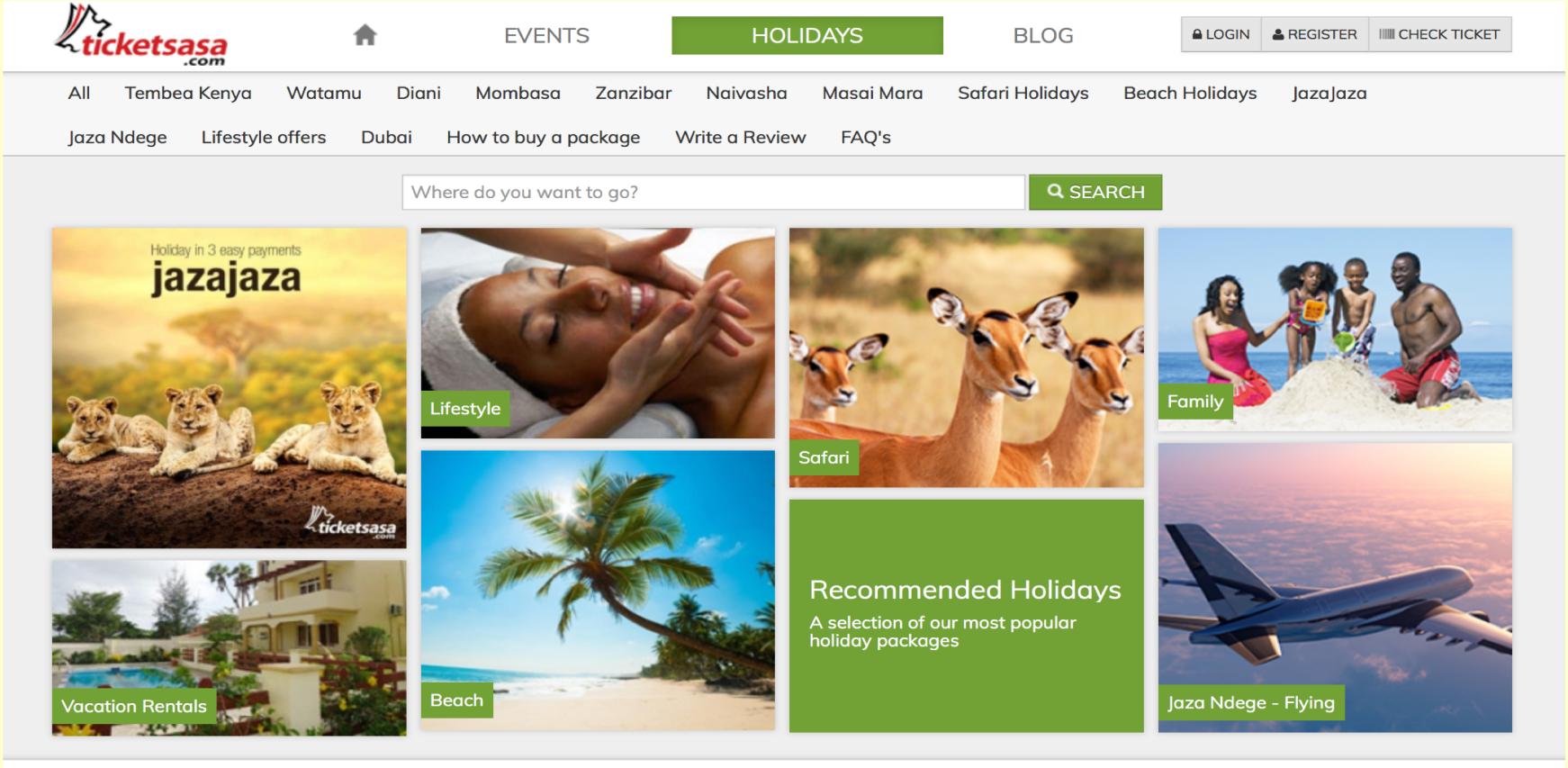
Webpage data extraction and text analytics: Ticketsasa

Organization question	Raw material (Data)	Data source	Data type
Can we get some insights from ticketsasa.com?	Yes	External webpage	Unstructured

- Data type: Unstructured
- Data source: External (Webpage)

Webpage data extraction and text analytics: Ticketsasa

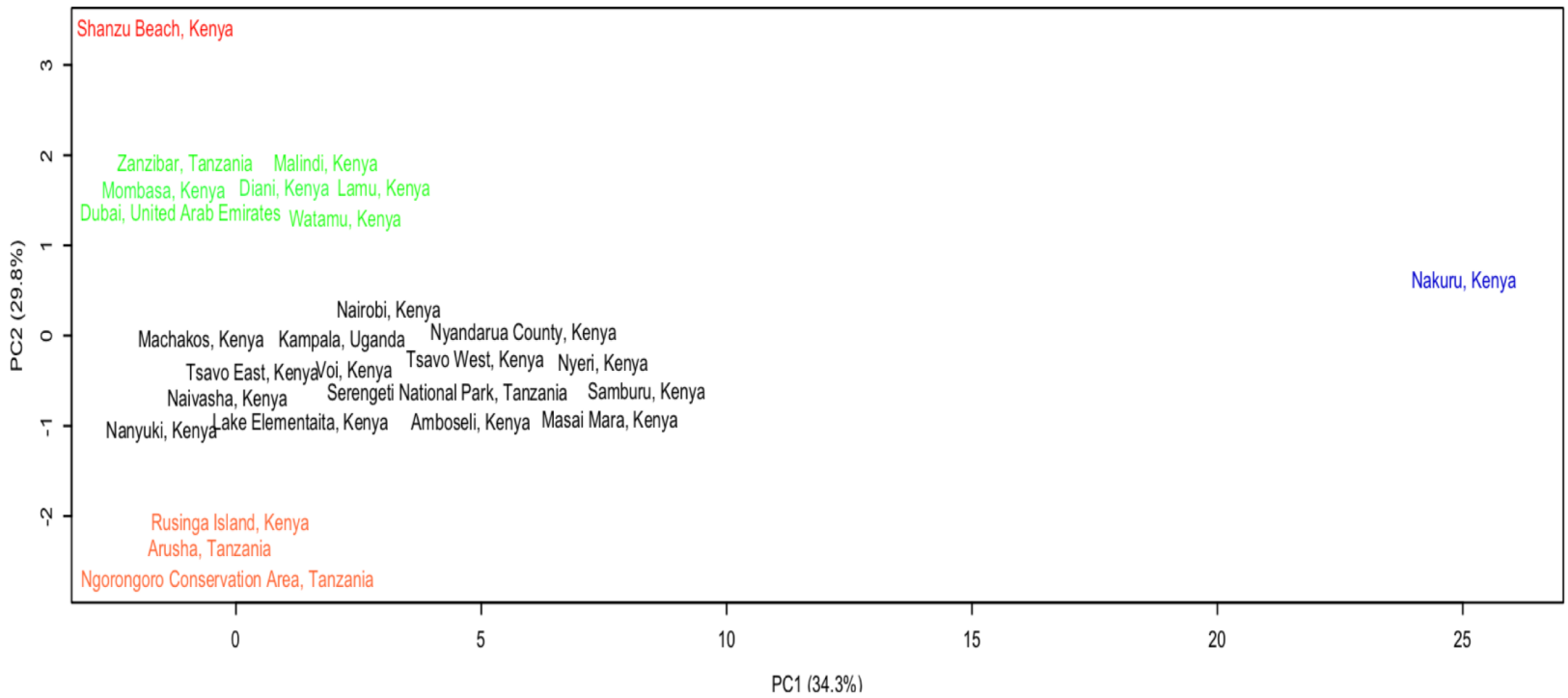
Organization question	Raw material (Data)	Data source	Data type
Can we get some insights from ticketsasa.com?	Yes	External webpage	Unstructured

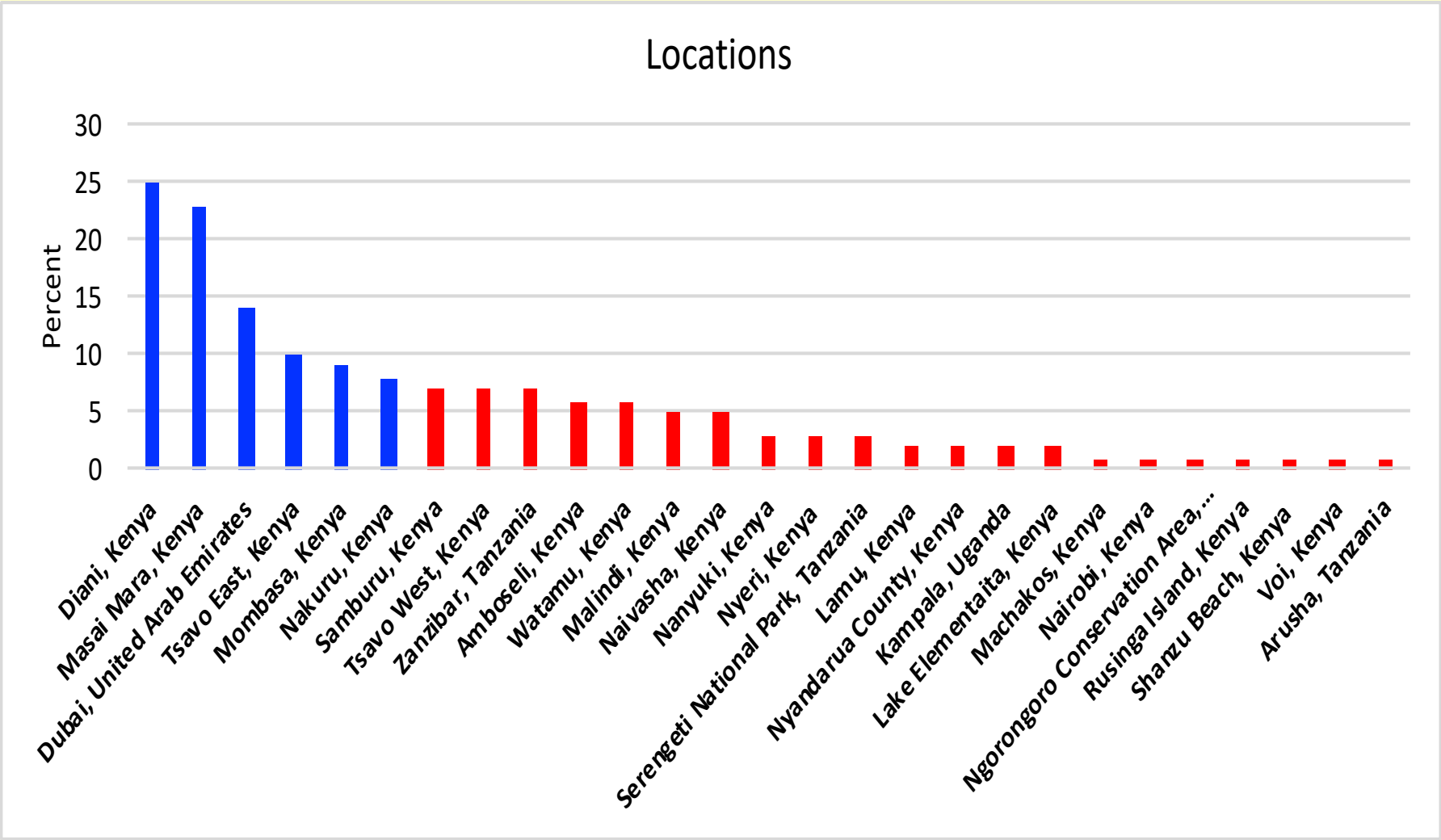


Scrapped data from Ticketsasa Holidays

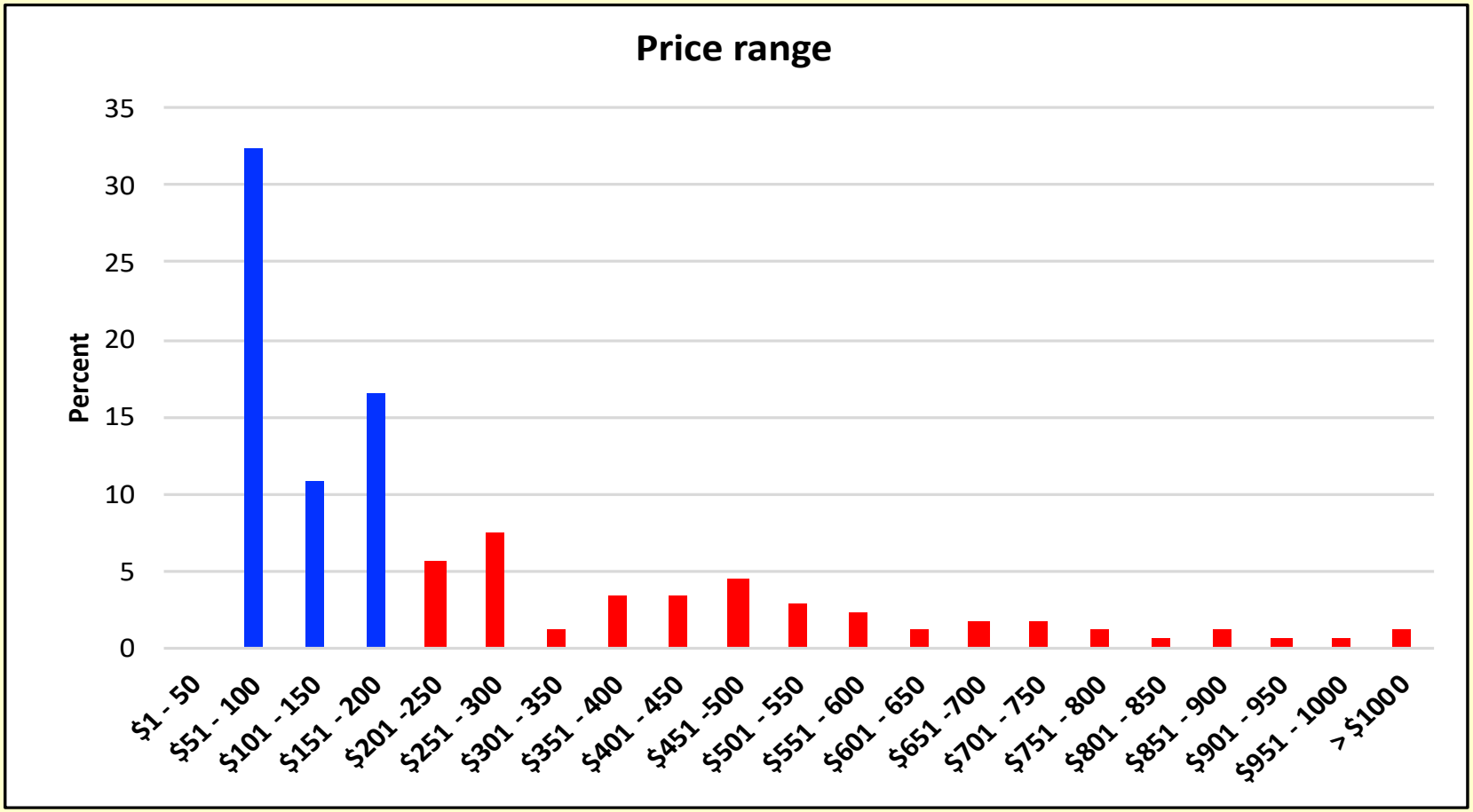
Description	Title	Url	Price	Includes	Location	Installinfo
Located on a private sanctuary on the western edge of Tsavo, the camp sits on a secluded dam on the Sante River ,Äì home to turtles, crocodiles and hundreds of basking hippos.	Safari Offer to Voyager Ziwani	https://www.ticketsasa.com/holidays/eventdetail/view/1858/safari_offer_to_voyager_ziwani	KES 6,100	Accommodation in a standard room (per night),Meals on full board basis	Tsavo West, Kenya	You can pay in 3 installments!
Tafaria Castle, sits on a hill overlooking Mou	Travel Offer to Tafaria Castle and (	https://www.ticketsasa.com/holidays/eve	KES 8,400	Accommodation (per night),M	Nyandarua County, Ker	You can pay in 3 installments!
Sarova Woodlands Hotel and Spa is a tranqu	Weekend Plan to Sarova Woodlanc	https://www.ticketsasa.com/holidays/eve	KES 9,450	Accommodation in a standarc	Nakuru, Kenya	You can pay in 3 installments!
On the shores of the Indian Ocean, at the e	Romantic Getaway to Driftwood B	https://www.ticketsasa.com/holidays/eve	KES 8,500	Accommodation in a standarc	Malindi, Kenya	You can pay in 3 installments!
Sarova Woodlands Hotel and Spa is a tranqu	Travel Offer to Sarova Woodlands	https://www.ticketsasa.com/holidays/eve	KES 18,399	2 nights accommodation,Mea	Nakuru, Kenya	You can pay in 3 installments!
Samburu Intrepids is a lush oasis on the bar	Luxury Safari to Samburu Intrepids	https://www.ticketsasa.com/holidays/eve	KES 7,495	Accommodation in a standarc	Samburu, Kenya	You can pay in 3 installments!
Ol Tukai Lodge,often referred to by its apt a	Luxury Safari to Ol Tukai Lodge	https://www.ticketsasa.com/holidays/eve	KES 10,050	Accommodation in a standarc	Amboseli, Kenya	You can pay in 3 installments!
Set amidst the turquoise Indian Ocean in th	Beach Getaway to Seven Islands R	https://www.ticketsasa.com/holidays/eve	KES 6,500	Accommodation (per night),M	Watamu, Kenya	You can pay in 3 installments!
Nestled in the rain forest, built right next to	Safari Offer to Shimba Hills Lodge	https://www.ticketsasa.com/holidays/eve	KES 12,899	Child up to three years sharin	Diani, Kenya	You can pay in 3 installments!
Kibali Villas Resort, your home away from h	Beach Holiday to Kibali Villas Reso	https://www.ticketsasa.com/holidays/eve	KES 7,000	Accommodation only	Malindi, Kenya	You can pay in 3 installments!
Mara River Lodge commands one of most s	Safari Offer to Mara River Lodge	https://www.ticketsasa.com/holidays/eve	KES 7,000	Accommodation in a standarc	Masai Mara, Kenya	You can pay in 3 installments!
Ilkeliani ¬†echo camp is located within the	Safari Offer at ilkeliani Camp	https://www.ticketsasa.com/holidays/eve	KES 8,500	Meals on full board basis.,Acc	Masai Mara, Kenya	You can pay in 3 installments!
Nestled in the heart of elephant country, M	Safari Getaway to Manyatta Camp	https://www.ticketsasa.com/holidays/eve	KES 7,500	Accommodation in a standarc	Tsavo East, Kenya	You can pay in 3 installments!
This superb complex is one of Kenya's most	All Inclusive holiday at Turtle Bay B	https://www.ticketsasa.com/holidays/eve	KES 7,800	Meals on all inclusive basis,A	Watamu, Kenya	You can pay in 3 installments!
This luxury tented camp is nestled on the ba	Travel Offer to Man-Eaters Lodge	https://www.ticketsasa.com/holidays/eve	KES 6,500	¬†Accommodation in a stand	Tsavo East, Kenya	You can pay in 3 installments!
The Leisure Lodge Beach & Golf Resort, wit	Family Vacation at Leisure Lodge G	https://www.ticketsasa.com/holidays/eve	KES 6,615	Accommodation in a Bahari s	Diani, Kenya	You can pay in 3 installments!
This luxury tented camp is nestled on the ba	Bush Retreat to Man-Eaters Lodge	https://www.ticketsasa.com/holidays/eve	KES 12,499	Children under the age of 4 st	Tsavo East, Kenya	You can pay in 3 installments!
Merely 260km say from Nairobi and nestled	3 Days 2 Nights at Ngulia Safari Lo	https://www.ticketsasa.com/holidays/eve	KES 9,499	2 nights accommodation,Mea	Tsavo West, Kenya	
Styled to echo the circular motif of a traditi	3 Day Flying Package to Mara Sere	https://www.ticketsasa.com/holidays/eve	KES 65,900	Children under 2 years stay fr	Masai Mara, Kenya	

Principal component plot using frequencies of words




Locations
Percent
30
25
20
15
10
5
0
Diani, Kenya
Masai Mara, Kenya
Dubai, United Arab Emirates
Tsavo East, Kenya
Mombasa, Kenya
Nakuru, Kenya
Samburu, Kenya
Tsavo West, Kenya
Zanzibar, Tanzania
Amboseli, Kenya
Watamu, Kenya
Malindi, Kenya
Naivasha, Kenya
Nanyuki, Kenya
Nyeri, Kenya
Serengeti National Park, Tanzania
Lamu, Kenya
Nyandarua County, Kenya
Kampala, Uganda
Lake Elementaita, Kenya
Machakos, Kenya
Nairobi, Kenya
Ngorongoro Conservation Area,...
Rusinga Island, Kenya
Shanzu Beach, Kenya
Voi, Kenya
Arusha, Tanzania

Price distributions



Analytics using church data: Easter attendance

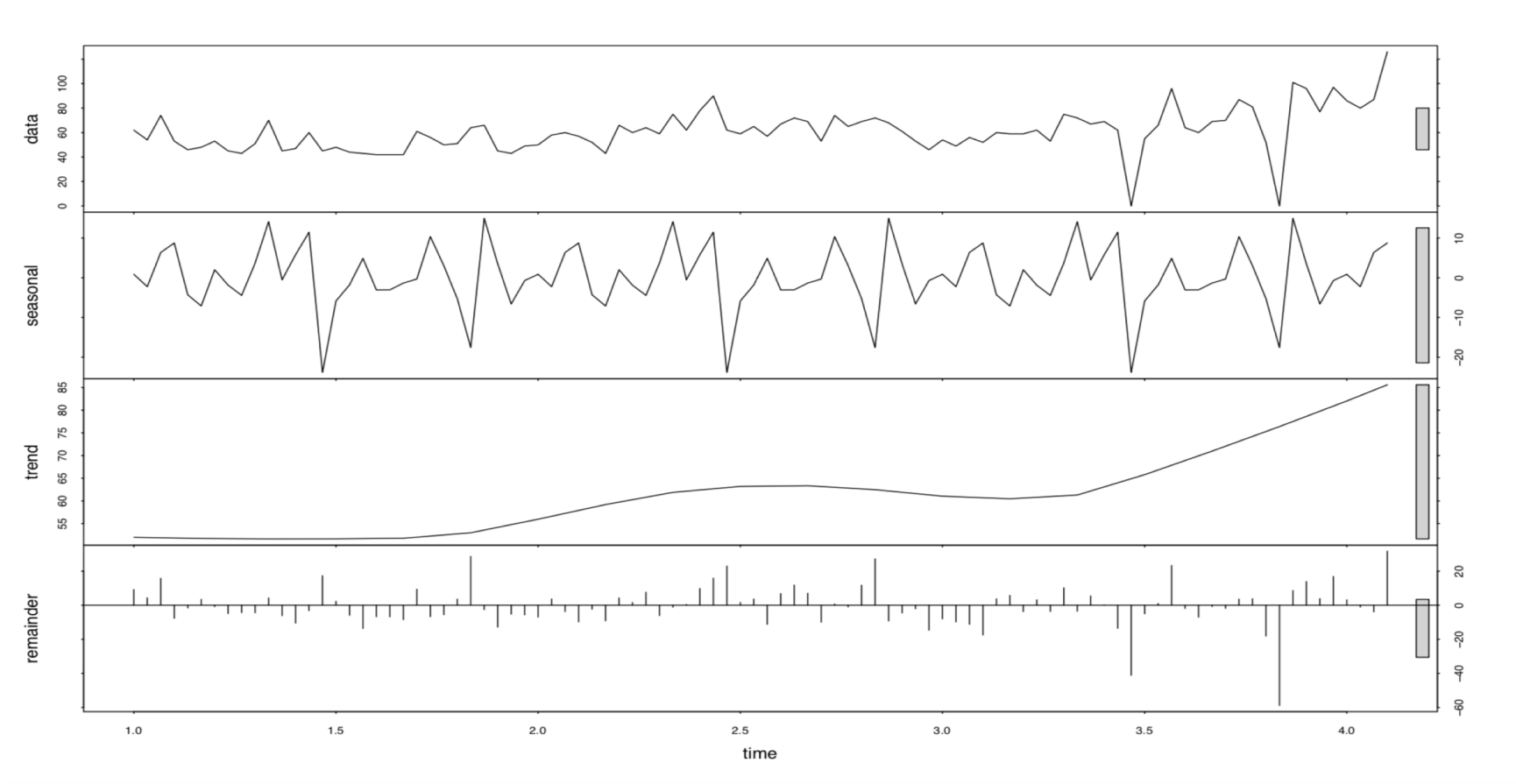
Organization question	Raw material (Data)	Data source	Data type
How many services do we need during Easter?	Yes	internal	Structured

- Format: Excel
- Data type: Structured
- Data source: Internal
- Observations: 6200
- Variables: 2

Sample data

Date	People
11/5/17	166
11/12/17	196
11/19/17	64
11/26/17	60
12/3/17	69
12/10/17	70
12/17/17	87
12/23/17	180
12/24/17	81
12/31/17	52
1/7/18	200
1/14/18	101
1/21/18	96
1/28/18	77
2/4/18	97
2/11/18	86
2/18/18	80
2/25/18	87
3/4/18	126

Graphs from Impact City Church



Results

- Predicted 798 people plus/minus 30
- Range (768 to 828)
- Need four services
- 812 people attended church on Easter Sunday

Acknowledgements

- Attendees
- Muthoni Wanyoike
- Emma Okere
- Nairobi Women in Machine Learning and Data Science
- Patira Data Science Team

Questions ?????