Assignment 6 Kaggle Competition

Nicholas Jacob and Yechang Qi

2024-09-05

Preparation and Modeling

First we examined the missing data. Recycling the coding from earlier in the semester, we get the following reports on data that is missing.

```
numericSummaryFinal %>% kable() #display table nicely when knitted
```

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
sessionId	70071	0	0.000	70071	100.000	4708198750205.951	1232288000046873000	4687300007946500000	4470700002792180000		7943023.750
custId	70071	0	0.000	47249	67.430	48874.987	1795	25081	48673	72587.50	96290.00	27321.879
visitStartTime	70071	0	0.000	69951	99.829	1485110879.478	1470355067	1477552704	1484102064	1493099921.50	1501655869.00	6581.658
visitNumber	70071	0	0.000	155	0.221	3.146	1	1	1	2.00	155.00	8.660
timeSinceLastVisit	70071	0	0.000	20970	29.927	256450.236	0	0	0	10375.00	30074517.00	1164717.354
isMobile	70071	0	0.000	2	0.003	0.229	0	0	0	0.00	1.00	0.420
isTrueDirect	70071	0	0.000	2	0.003	0.400	0	0	0	1.00	1.00	0.490
adwordsClickInfo.page	70071	68260	97.415	6	0.009	1.008	1	1	1	1.00	7.00	0.179
pageviews	70071	8	0.011	155	0.221	6.304	1	1	2	6.00	469.00	11.693
bounces	70071	40719	58.111	2	0.003	1.000	1	1	1	1.00	1.00	0.000
newVisits	70071	23944	34.171	2	0.003	1.000	1	1	1	1.00	1.00	0.000
revenue	70071	0	0.000	5850	8.349	10.165	0	0	0	0.00	15980.79	99.534
totalRevenue	70071	0	0.000	5085	7.257	69.006	0	0	0	9.98	15980.79	442.820
targetRevenue	70071	0	0.000	5085	7.257	1.132	0	0	0	2.40	9.68	2.037

```
options(digits=3)
options(scipen=99)
factorSummaryFinal %>% kable() #display nicely
```

variable	n	missing	missing_pct	unique	unique_pct	freq_Ratio	mode	first_mode_freq	2nd mode	second_mode_freq	least common	least_common_freq	least common freq
date	70071	0	0.000	366	0.522	1.03	2016-12-05	362	2016-11-28	352	2017-05-28		76
channelGrouping	70071	0	0.000	8	0.011	2.03	Organic Search	27503	Social	13528	(Other)		3
browser	70071	0	0.000	28	0.040	4.30	Chrome	51584	Safari	12007	Apple-iPhone7C2		1
operatingSystem	70071	307	0.438	16	0.023	1.01	Macintosh	23970	Windows	23707	Nintendo 3DS		1
deviceCategory	70071	0	0.000	3	0.004	3.89	desktop	53986	mobile	13868	tablet		2217
continent	70071	85	0.121	6	0.009	3.10	Americas	42508	Asia	13697	Africa		888
subContinent	70071	85	0.121	23	0.033	8.06	Northern America	38860	Southeast Asia	4823	Polynesia		1

variable	n	missing	missing_pct	unique	unique_pct	ratio	mode	freq	2nd common mode	freq	least common	least common freq
country	70078	85	0.12	1177	0.25	12.14	United States	36941	India	3044	Anguilla	1
region	70078	38485	54.92	310	0.44	3.25	California	11254	New York	3468	Abruzzo	1
metro	70078	49183	70.19	73	0.10	2.86	San Francisco-Oakland-San Jose CA	10072	New York NY	3526	Albany-Schenectady-Troy NY	1
city	70078	39002	55.69	878	0.68	1.32	Mountain View	4569	New York	3465	Adelaide	1
networkDomain	70078	61334	87.73	15015	7.15	2.11	comcast.net	2890	verizon.net	1372	0rbitel.net	1
topLevelDomain	70078	61334	87.73	84	0.26	2.39	net	15027	com	6297	25	1
campaign	70078	67319	96.06	70	0.01	1.35	AW - Dynamic Search Ads Whole Site	1229	Data Share Promo	911	Retail (DO NOT EDIT owners nophakun and tianyu)	4
source	70078	2	0.00	1132	0.18	2.30	google	29233	youtube.com	12708	0.shared.bow.cat2.ads-bow.lf.borg.google.com:9860	1
medium	70078	11827	16.87	9	0.00	1.02	organic	27503	referral	27010	cpm	735
keyword	70078	67419	96.20	516	0.59	4.68	6qEhsCssdK0z3	997	1hZbAqLCbjwfg2	217	+Google +accessories	1
referralPath	70078	43006	61.45	384	0.54	2.62	/	11419	/yt/about/	4359	//scs/mail-static//js/k=gmail.main.en.PDqUrNqDME8.O/m=v-8PxjUMA7rSV-vO-_7ne0v9h37u8f_JgKjkVaD_zf4fxP-B_LSNAg/rt=h/d=1/rs=AHGWq9BLPEHaiYtr	1
adContent	70078	69239	98.80	28	0.04	5.48	Google Merchandise Collection	449	Display Ad created 3/11/14	82	{KeyWord:Want Google Pet Toys?}	1
adwordsClickInfo.slot	70078	68260	97.45	5	0.00	44.27	Top	1771	RHS	40	RHS	40
adwordsClickInfo	70078	68265	97.39	4406	2.00	1.40	Cj0KEQiA6_TBBRDi...	...	...	...	CWitE0IY2pK5BQqdZzoJG2eW4TQQWxthoCuI_...	1
adwordsClickInfo.adNetworkType	70078	68260	97.41	2	0.00	1.00	Google Search	1811	NA	-Inf	Google Search	1811