CAPSTONE PROJECT

VODKA DATA

K VINEET PATNAIK

INTRODUCTION

The Dataset given was related to Vodka industry data with different variables for sales , quantities and Brands etc.It is related to mix marketing model and we have to work out questions based on the data and analyze as a business model solution .

Most of the variables are related to :

(i) Marketing variables

(ii) Sales numbers

(iii) Brands

(iv) Year

TOOLS USED & STEPS FOLLOWED

1. We have used MS-Excel to read the data and understand the data .

2. Data Cleaning and complete analysis done on Python
   - Info and summary (some datatypes changed)
   - Removed null values(NA)
   - Basic analysis (Graphs and statistical analysis)
   - Advanced ML models deployed (Regression analysis)

BASIC ANALYSIS

- EDA of the data was done
- Boxplots for the numerical data is plotted and outliers are found
- Plots for Sales ,ads and Revenue generated were plotted w.r.t Year
  - With increase in ad amount ,Sales have been increased w.r.t time
- Plots for Sales ,ads and Revenue generated were plotted w.r.t Brands
  - Most of the Sales have been increased w.r.t ad money but for few even with less money spent on ads they had good revenues .
- Biggest Brands in terms of Dollar Sales and price per unit
- Tier-1 & Tier-2 average and total sales were analyzed
- Domestic and International Brands Sales were analyzed

STATISTICAL ANALYSIS

	Year	TotalSales	LagTotalSales	2LagTotalSales	LnSales	LnLSales	Ln2Lsales	LnDiff	diff	IfDom	DollarSales	PriceRerUnit	LagPrice	LnPrice
count	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000
mean	2001.776000	1438.668000	1389.352000	1346.016000	6.758185	6.701590	6.631309	0.056065	0.076997	0.624000	115330.803200	82.218204	233.196715	4.202910
std	3.610258	1618.046372	1564.697629	1524.277431	1.076482	1.117083	1.189156	0.175100	0.247650	0.485352	165282.931365	55.920256	1721.078091	0.628922
min	1995.000000	47.000000	27.000000	18.000000	3.850147	3.295837	2.890372	-0.563493	-0.396364	0.000000	5320.000000	27.291230	27.291230	3.306566
25%	1999.000000	515.000000	485.000000	436.250000	6.244027	6.184149	6.078203	-0.019994	-0.019796	0.000000	24393.500000	37.825868	37.762208	3.632992
50%	2002.000000	921.000000	886.500000	859.500000	6.825455	6.787128	6.756349	0.034324	0.034920	1.000000	44711.500000	52.955750	50.910035	3.969453
75%	2005.000000	1735.500000	1647.500000	1592.750000	7.459015	7.407011	7.373217	0.095610	0.100330	1.000000	104000.000000	122.000000	117.321700	4.804021
max	2007.000000	9015.000000	8505.000000	8149.000000	9.106646	9.048409	9.005651	1.146814	2.148148	1.000000	786721.000000	250.262300	20806.000000	5.522510

LnLPrice	Mag	News	Outdoor	Broad	Print	LnMag	LnNews	LnOut	LnBroad	LnPrint	Tier1	Tier2	TotalMinusSales	LagTotalMinu
250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250
4.223238	3337.648800	193.822800	169.244400	282.123600	3531.471596	3.036801	1.532045	1.397279	1.001880	3.346910	0.228000	0.236000	62618.164000	62635
0.793282	6770.112453	557.300384	477.872973	1063.072155	6954.582102	4.159412	2.666546	2.560691	2.450855	4.242027	0.420384	0.425474	1622.532606	1571
3.306566	1.000000	0.300000	1.000000	0.600000	2.000000	0.000000	-1.203973	0.000010	-0.510826	0.000010	0.000000	0.000000	55209.000000	55687
3.631309	1.000000	1.000000	1.000000	1.000000	2.000000	0.000000	0.000000	0.000010	0.000010	0.000010	0.000000	0.000000	62409.750000	62439
3.929355	1.000000	1.000000	1.000000	1.000000	2.000000	0.000000	0.000000	0.000010	0.000010	0.000010	0.000000	0.000000	63054.500000	63103
4.764920	4033.750000	41.775000	20.125000	1.000000	4197.775000	8.136850	3.385083	0.000010	0.000010	8.342153	0.000000	0.000000	63599.250000	63599
9.942997	33971.300000	3524.900000	3255.400000	7827.100000	34504.700000	10.433270	8.167607	8.088071	8.965347	10.448850	1.000000	1.000000	64163.000000	64131

LagTotalMinusSales	TierSales	OutsideTierSales	LagTierSales	LagOutsideTierSales	Firstintro	Marketshare	LagMktshare	YearID	total ad
250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000
62635.480000	9562.796000	53055.368000	9237.680000	53397.800000	0.016000	0.048302	0.048947	9.776000	7514.311196
1571.495717	2927.199324	2208.939172	2960.213039	2282.303824	0.125727	0.054646	0.055499	3.610258	14432.952198
55687.000000	846.000000	48358.000000	697.000000	49806.000000	0.000000	0.001468	0.000971	3.000000	6.000000
62439.000000	8371.250000	51863.000000	7662.750000	51947.000000	0.000000	0.015354	0.015109	7.000000	6.000000
63103.000000	10670.000000	52335.000000	10440.000000	52419.000000	0.000000	0.033718	0.033718	10.000000	6.000000
63599.250000	11181.750000	53806.000000	11125.250000	54669.000000	0.000000	0.053621	0.053754	13.000000	10250.625000
64131.000000	15790.000000	59760.000000	14299.000000	59868.000000	1.000000	0.270477	0.270477	15.000000	70489.200000

REGRESSION ANALYSIS

Q1 Run a regression of the natural logarithm of sales on all the following price: price, print, marketing expenditure, outdoor marketing expenditure, broadcast marketing expenditure, and previous years sale. Evaluate the results. Perform residual analysis to satisfy the assumptions of regression.

R-squared value - 0.774
Which means these variables had explained the model 77.4

Remaining part explained and worked in Google.colab

OLS Regression Results

Dep. Variable:	LnSales	R-squared:	0.774
Model:	OLS	Adj. R-squared:	0.769
Method:	Least Squares	F-statistic:	166.7
Date:	Sat, 20 Feb 2021	Prob (F-statistic):	1.44e-76
Time:	13:34:28	Log-Likelihood:	-187.01
No. Observations:	250	AIC:	386.0
Df Residuals:	244	BIC:	407.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5096	0.084	77.587	0.000	6.344	6.675
Print	4.889e-05	7.28e-06	6.715	0.000	3.46e-05	6.32e-05
PriceRerUnit	-0.0070	0.001	-9.003	0.000	-0.009	-0.005
Outdoor	-0.0003	9.39e-05	-3.508	0.001	-0.001	-0.000
Broad	-4.246e-05	4.58e-05	-0.926	0.355	-0.000	4.78e-05
LagTotalSales	0.0005	3.71e-05	13.932	0.000	0.000	0.001

Omnibus:	16.547	Durbin-Watson:	0.660
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.328
Skew:	-0.578	Prob(JB):	0.000105
Kurtosis:	3.649	Cond. No.	2.03e+04

Q2. Run a regression of the natural logarithm of change in sales on the natural logarithm of previous period's prices and the natural log of marketing expenditures on print, outdoor and broadcasting.Evaluate the results. Perform residual analysis to satisfy the assumptions of regression

In this model ,Natural log of previous period's prices is the variable that has highest effect wth coeff (0.85)

Broadcasting has the highest impact on sales in terms of marketing (coeff = 0.15)

OLS Regression Results

Dep. Variable:	LnSales	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.415
Method:	Least Squares	F-statistic:	45.10
Date:	Sat, 20 Feb 2021	Prob (F-statistic):	2.34e-28
Time:	12:40:22	Log-Likelihood:	-303.69
No. Observations:	250	AIC:	617.4
Df Residuals:	245	BIC:	635.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.8624	0.336	29.350	0.000	9.200	10.524
LnLPrice	-0.8513	0.086	-9.885	0.000	-1.021	-0.682
LnPrint	0.0657	0.021	3.197	0.002	0.025	0.106
LnOut	0.0843	0.029	2.903	0.004	0.027	0.141
LnBroad	0.1530	0.026	5.912	0.000	0.102	0.204

Omnibus:	6.645	Durbin-Watson:	0.498
Prob(Omnibus):	0.036	Jarque-Bera (JB):	6.377
Skew:	-0.357	Prob(JB):	0.0412
Kurtosis:	3.322	Cond. No.	46.7

3a. To understand the influence of Vodka quality, run a regression by adding the tier 1 and tier 2 dummy variables (that indicate whether a vodka brand belongs to first or second quality tiers) to the set of independent variables in Q2. Evaluate the results. Perform residual analysis to satisfy the assumptions of regression

Here Tier-1 & 2 have the highest coeff value , so high impact than any other value, so quality of vodka plays a significant role

Brands were dummied and some of the brand had huge impact.

R-squared - 0.133

OLS Regression Results

Dep. Variable:	LnDiff	R-squared:	0.133
Model:	OLS	Adj. R-squared:	0.111
Method:	Least Squares	F-statistic:	6.207
Date:	Sun, 21 Feb 2021	Prob (F-statistic):	4.44e-06
Time:	06:32:16	Log-Likelihood:	99.191
No. Observations:	250	AIC:	-184.4
Df Residuals:	243	BIC:	-159.7
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1430	0.083	1.727	0.085	-0.020	0.306
Tier1[T.1]	0.1450	0.058	2.522	0.012	0.032	0.258
Tier2[T.1]	0.1305	0.044	2.996	0.003	0.045	0.216
LnLPrice	-0.0377	0.022	-1.695	0.091	-0.081	0.006
LnPrint	0.0093	0.005	1.835	0.068	-0.001	0.019
LnOut	-0.0128	0.006	-2.106	0.036	-0.025	-0.001
LnBroad	-0.0049	0.005	-0.923	0.357	-0.015	0.006

Omnibus:	118.599	Durbin-Watson:	1.331
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1574.191
Skew:	1.502	Prob(JB):	0.00
Kurtosis:	14.920	Cond. No.	65.9

b.

Brands were dummied and some of the brand had huge impact.

R-squared - 0.563

Intercept	0.3783	0.330	1.146	0.253	-0.272	1.029
Tier1[T.1]	0.1459	0.131	1.113	0.267	-0.112	0.404
Tier2[T.1]	0.1896	0.085	2.243	0.026	0.023	0.356
BrandName[T.Aristocrat]	-0.0014	0.034	-0.041	0.967	-0.068	0.065
BrandName[T.Barton]	0.0105	0.033	0.312	0.755	-0.056	0.076
BrandName[T.Belvedere]	0.1855	0.089	2.093	0.037	0.011	0.360
BrandName[T.Burnett]	0.1084	0.043	2.533	0.012	0.024	0.193
BrandName[T.Chopin]	0.1338	0.097	1.377	0.170	-0.058	0.325
BrandName[T.Crystal Palac]	0.0071	0.033	0.211	0.833	-0.059	0.073
BrandName[T.Finlandia]	-0.0526	0.053	-0.987	0.325	-0.158	0.052
BrandName[T.Fleischmann's]	0.0182	0.039	0.472	0.638	-0.058	0.094
BrandName[T.Fris]	-0.0174	0.057	-0.307	0.759	-0.129	0.094
BrandName[T.Gilbey's]	-0.0455	0.039	-1.164	0.246	-0.123	0.032
BrandName[T.Gordon's]	0.0057	0.053	0.108	0.914	-0.099	0.110
BrandName[T.Grey Goose]	0.6332	0.073	8.717	0.000	0.490	0.776
BrandName[T.Kamchatka]	-0.0334	0.034	-0.971	0.332	-0.101	0.034
BrandName[T.Ketel One]	0.1201	0.067	1.790	0.075	-0.012	0.252
BrandName[T.Level]	0.5216	0.487	1.071	0.285	-0.438	1.482
BrandName[T.McCormick]	-0.0009	0.040	-0.022	0.982	-0.079	0.077
BrandName[T.Polar Ice]	-0.0718	0.082	-0.874	0.383	-0.234	0.090
BrandName[T.Popov]	-0.0535	0.037	-1.444	0.150	-0.127	0.020
BrandName[T.Pravda]	-0.1511	0.098	-1.540	0.125	-0.344	0.042
BrandName[T.Skol]	0.0277	0.035	0.797	0.426	-0.041	0.096
BrandName[T.Sky]	-0.0086	0.049	-0.174	0.862	-0.105	0.088
BrandName[T.Smirnoff]	-0.0914	0.048	-1.909	0.058	-0.186	0.003
BrandName[T.Stolicnaya]	0.0863	0.054	1.612	0.108	-0.019	0.192
BrandName[T.Tanqueray]	-0.0703	0.051	-1.389	0.166	-0.170	0.029
BrandName[T.Three Olives]	0.5016	0.050	10.019	0.000	0.403	0.600
LnLPrice	-0.1019	0.093	-1.094	0.275	-0.286	0.082
LnPrint	-0.0093	0.005	-1.936	0.054	-0.019	0.000
LnOut	0.0240	0.006	3.998	0.000	0.012	0.036
LnBroad	-0.0125	0.005	-2.369	0.019	-0.023	-0.002

4a. To understand the influence of competition and brand power, run a regression by adding the sum of sales of all the competing brands in the previous year ("lagtotalminussales") to the independent variables in Q3. Perform residual analysis to satisfy the assumptions of regression.

LagTotalMinusSales had very less impact.Much lesser than Tier-1 or 2 .Although the model has improved by a small margin

R-squared - 0.177

OLS Regression Results

Dep. Variable:	LnDiff	R-squared:	0.177
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	7.445
Date:	Sun, 21 Feb 2021	Prob (F-statistic):	4.12e-08
Time:	06:22:14	Log-Likelihood:	105.75
No. Observations:	250	AIC:	-195.5
Df Residuals:	242	BIC:	-167.3
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.7876	0.541	-3.305	0.001	-2.853	-0.722
Tier1[T.1]	0.1293	0.056	2.296	0.023	0.018	0.240
Tier2[T.1]	0.1211	0.043	2.843	0.005	0.037	0.205
LnLPrice	-0.0572	0.022	-2.557	0.011	-0.101	-0.013
LnPrint	0.0115	0.005	2.306	0.022	0.002	0.021
LnOut	-0.0060	0.006	-0.962	0.337	-0.018	0.006
LnBroad	0.0036	0.006	0.621	0.535	-0.008	0.015
LagTotalMinusSales	3.183e-05	8.82e-06	3.609	0.000	1.45e-05	4.92e-05

Omnibus:	118.672	Durbin-Watson:	1.333
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1804.653
Skew:	1.455	Prob(JB):	0.00
Kurtosis:	15.837	Cond. No.	3.33e+06

b

If Brands are dummied ,then R-squared value increases by a huge margin

R-squared - 0.599

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.4157	1.493	-4.298	0.000	-9.358	-3.474
Tier1[T.1]	-0.1474	0.140	-1.051	0.294	-0.424	0.129
Tier2[T.1]	-0.3081	0.134	-2.299	0.022	-0.572	-0.044
BrandName[T.Aristocrat]	-0.5961	0.132	-4.526	0.000	-0.856	-0.337
BrandName[T.Barton]	-0.4981	0.114	-4.376	0.000	-0.722	-0.274
BrandName[T.Belvedere]	-0.2688	0.129	-2.080	0.039	-0.523	-0.014
BrandName[T.Burnett]	-0.5218	0.141	-3.690	0.000	-0.800	-0.243
BrandName[T.Chopin]	-0.3166	0.134	-2.361	0.019	-0.581	-0.052
BrandName[T.Crystal Palac]	-0.5948	0.133	-4.467	0.000	-0.857	-0.332
BrandName[T.Finlandia]	-0.2518	0.067	-3.786	0.000	-0.383	-0.121
BrandName[T.Fleischmann's]	-0.5733	0.132	-4.334	0.000	-0.834	-0.313
BrandName[T.Fris]	-0.2219	0.070	-3.178	0.002	-0.359	-0.084
BrandName[T.Gilbey's]	-0.6366	0.132	-4.810	0.000	-0.897	-0.376
BrandName[T.Gordon's]	-0.4776	0.115	-4.136	0.000	-0.705	-0.250
BrandName[T.Grey Goose]	0.2624	0.106	2.484	0.014	0.054	0.471
BrandName[T.Kamchatka]	-0.5946	0.125	-4.759	0.000	-0.841	-0.348
BrandName[T.Ketel One]	-0.1728	0.090	-1.923	0.056	-0.350	0.004
BrandName[T.Level]	-0.0978	0.484	-0.202	0.840	-1.052	0.857
BrandName[T.McCormick]	-0.4783	0.109	-4.377	0.000	-0.694	-0.263
BrandName[T.Polar Ice]	-0.2795	0.090	-3.094	0.002	-0.458	-0.101
BrandName[T.Popov]	-0.4740	0.097	-4.886	0.000	-0.665	-0.283
BrandName[T.Pravda]	-0.5802	0.131	-4.413	0.000	-0.839	-0.321
BrandName[T.Skol]	-0.5150	0.121	-4.248	0.000	-0.754	-0.276
BrandName[T.Sky]	-0.0500	0.048	-1.046	0.297	-0.144	0.044
BrandName[T.Smirnoff]	0.4240	0.120	3.540	0.000	0.188	0.660
BrandName[T.Stolicnaya]	-0.1965	0.079	-2.475	0.014	-0.353	-0.040
BrandName[T.Tanqueray]	-0.2564	0.063	-4.086	0.000	-0.380	-0.133
BrandName[T.Three Olives]	0.3274	0.061	5.389	0.000	0.208	0.447
LnLPrice	-0.0632	0.090	-0.706	0.481	-0.240	0.113
LnPrint	-0.0101	0.005	-2.190	0.030	-0.019	-0.001
LnOut	0.0244	0.006	4.253	0.000	0.013	0.036
LnBroad	-0.0026	0.005	-0.478	0.633	-0.013	0.008
LagTotalMinusSales	0.0001	2.46e-05	4.656	0.000	6.6e-05	0.000

5. To measure the sales growth of new brands compared to the existent ones, include the variable “firstintro” to the independent variable set in Q4. Firstintro is equal to one in the first three years after a brand is introduced and is zero elsewhere

Firstintro had the most impact(coeff - 0.54) than Tier-1 & 2 and LagtotalminusSales and the model has improved

R-squared - 0.314

OLS Regression Results

Dep. Variable:	LnDiff	R-squared:	0.314
Model:	OLS	Adj. R-squared:	0.292
Method:	Least Squares	F-statistic:	13.81
Date:	Sun, 21 Feb 2021	Prob (F-statistic):	1.79e-16
Time:	06:19:47	Log-Likelihood:	128.54
No. Observations:	250	AIC:	-239.1
Df Residuals:	241	BIC:	-207.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.6243	0.495	-3.279	0.001	-2.600	-0.648
Tier1[T.1]	0.1273	0.052	2.471	0.014	0.026	0.229
Tier2[T.1]	0.1287	0.039	3.303	0.001	0.052	0.205
LnLPrice	-0.0865	0.021	-4.139	0.000	-0.128	-0.045
LnPrint	0.0129	0.005	2.824	0.005	0.004	0.022
LnOut	-0.0079	0.006	-1.380	0.169	-0.019	0.003
LnBroad	0.0076	0.005	1.440	0.151	-0.003	0.018
LagTotalMinusSales	3.093e-05	8.07e-06	3.835	0.000	1.5e-05	4.68e-05
Firstintro	0.5427	0.078	6.943	0.000	0.389	0.697

b

If Brands are dummied ,then R-squared value increases by a huge margin

R-squared - 0.669

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.7928	1.362	-4.987	0.000	-9.477	-4.108
Tier1[T.1]	-0.4570	0.136	-3.364	0.001	-0.725	-0.189
Tier2[T.1]	-0.4974	0.125	-3.969	0.000	-0.744	-0.250
BrandName[T.Aristocrat]	-0.5610	0.120	-4.668	0.000	-0.798	-0.324
BrandName[T.Barton]	-0.4624	0.104	-4.452	0.000	-0.667	-0.258
BrandName[T.Belvedere]	-0.3797	0.119	-3.193	0.002	-0.614	-0.145
BrandName[T.Burnett]	-0.5245	0.129	-4.069	0.000	-0.779	-0.270
BrandName[T.Chopin]	-0.4091	0.123	-3.325	0.001	-0.652	-0.167
BrandName[T.Crystal Palac]	-0.5495	0.122	-4.520	0.000	-0.789	-0.310
BrandName[T.Finlandia]	-0.3359	0.062	-5.427	0.000	-0.458	-0.214
BrandName[T.Fleischmann's]	-0.5625	0.121	-4.665	0.000	-0.800	-0.325
BrandName[T.Fris]	-0.2761	0.064	-4.305	0.000	-0.403	-0.150
BrandName[T.Gilbey's]	-0.6417	0.121	-5.319	0.000	-0.879	-0.404
BrandName[T.Gordon's]	-0.5463	0.106	-5.166	0.000	-0.755	-0.338
BrandName[T.Grey Goose]	0.1364	0.098	1.390	0.166	-0.057	0.330
BrandName[T.Kamchatka]	-0.5729	0.114	-5.028	0.000	-0.797	-0.348
BrandName[T.Ketel One]	-0.1429	0.082	-1.741	0.083	-0.305	0.019
BrandName[T.Level]	-1.4138	0.483	-2.929	0.004	-2.365	-0.463
BrandName[T.McCormick]	-0.4406	0.100	-4.415	0.000	-0.637	-0.244
BrandName[T.Polar Ice]	-0.1962	0.083	-2.355	0.019	-0.360	-0.032
BrandName[T.Popov]	-0.4790	0.088	-5.416	0.000	-0.653	-0.305
BrandName[T.Pravda]	-0.8241	0.125	-6.582	0.000	-1.071	-0.577
BrandName[T.Skol]	-0.4978	0.111	-4.504	0.000	-0.716	-0.280
BrandName[T.Sky]	-0.0904	0.044	-2.056	0.041	-0.177	-0.004
BrandName[T.Smirnoff]	0.4249	0.109	3.891	0.000	0.210	0.640
BrandName[T.Stolicnaya]	-0.1851	0.072	-2.556	0.011	-0.328	-0.042
BrandName[T.Tanqueray]	-0.2722	0.057	-4.756	0.000	-0.385	-0.159
BrandName[T.Three Olives]	0.2485	0.057	4.391	0.000	0.137	0.360
LnLPrice	0.1529	0.088	1.745	0.082	-0.020	0.326
LnPrint	-0.0018	0.004	-0.401	0.689	-0.010	0.007
LnOut	0.0194	0.005	3.677	0.000	0.009	0.030
LnBroad	-0.0030	0.005	-0.612	0.541	-0.013	0.007
LagTotalMinusSales	0.0001	2.24e-05	4.800	0.000	6.34e-05	0.000
Firstintro	0.4922	0.073	6.749	0.000	0.348	0.636