MSCI -719 Operations Analytics

Names: Hemaseshan Rajasekaran (20800259) and Sushant Kataria (20771302)

Q1.A. The uncertainties that would affect the daily surgical volume are-

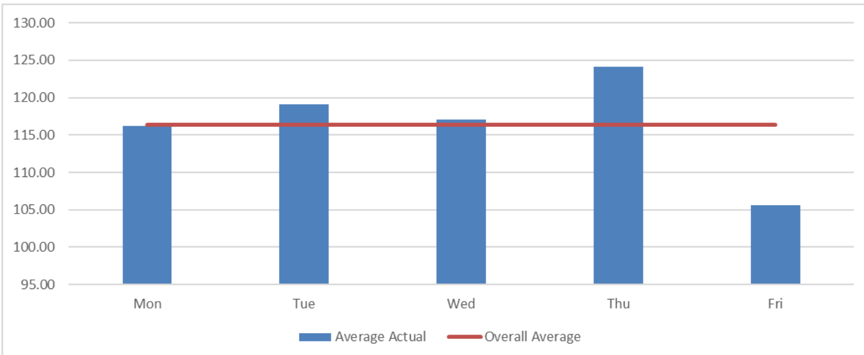
- Emergency cases will vary day by day.
- Surgeon's meeting or conferences or vacation days leads to low volume.
- Patient's needs and their own scheduling preferences (fewer emergency cases).
- National holidays/festivals like Christmas season see low demand.
- Surgeon's preference about scheduling either in beginning or end of week.

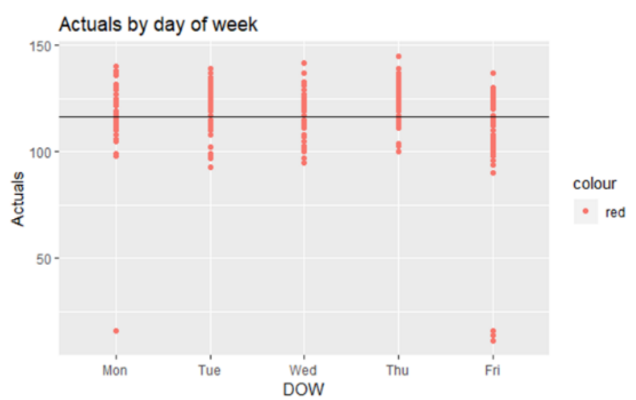
Q1.B By assessing daily case volume correctly, the ideal number of operating rooms(OR) required can be decided. This will help in reducing the total OR costs drastically. For every planned OR which was not used, the associated costs to keep it emergency ready is high and should still be paid. This includes fixed costs like electricity and oxygen supply. Labor costs for surgeons and nurses must be paid by the clock. Also, ancillary services like pathology, radiology and recovery room must also be paid for unoccupied OR. If more surgeries than planned were to happen, procuring supplies at the last minute will also incur additional costs.

Q2.A EDA to understand the elements in data

Element 1: day of the week

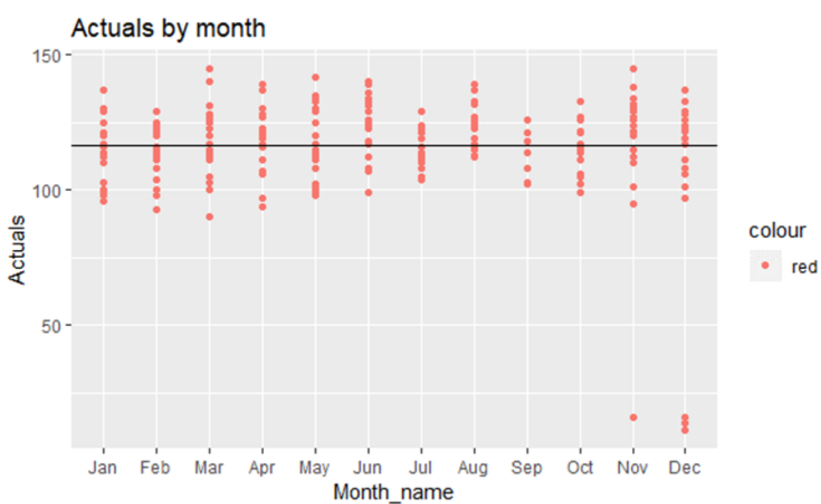
DOW	Average Actual	Overall Average
Mon	116.26	116.38
Tue	119.08	116.38
Wed	117.04	116.38
Thu	124.08	116.38
Fri	105.61	116.38





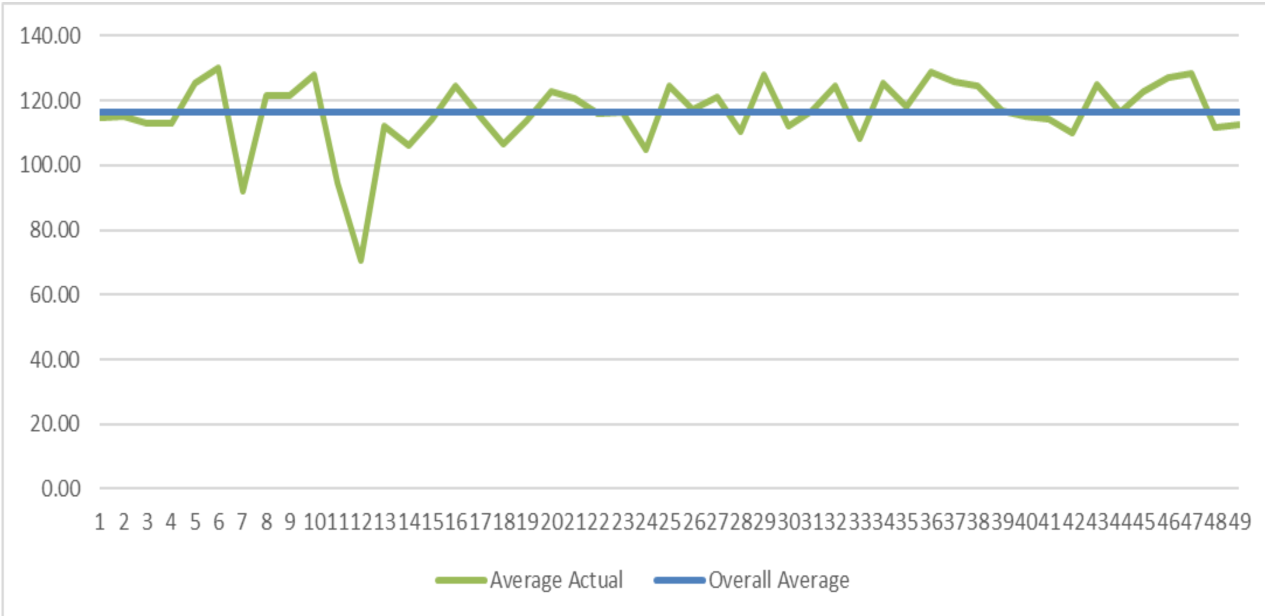
Element 2: Month

Month Nbr	Average Actual	Overall Average
1	114.77	116.38
2	114.33	116.38
3	116.50	116.38
4	118.57	116.38
5	117.59	116.38
6	124.10	116.38
7	114.48	116.38
8	124.35	116.38
9	112.11	116.38
10	113.69	116.38
11	118.19	116.38
12	104.64	116.38



Element 3: Week

Week Nbr	Average Actual	Overall Average
1	114.60	116.38
2	115.00	116.38
3	113.20	116.38
4	113.00	116.38
5	125.20	116.38
6	130.20	116.38
7	92.00	116.38
8	121.40	116.38
9	121.40	116.38
10	127.80	116.38
11	94.40	116.38
12	70.80	116.38
13	112.00	116.38
14	106.20	116.38
15	114.20	116.38
16	124.40	116.38
17	115.20	116.38
18	106.60	116.38
19	113.80	116.38
20	123.00	116.38
21	120.60	116.38
22	116.00	116.38
23	116.40	116.38
24	104.80	116.38
25	124.60	116.38
26	117.20	116.38
27	121.00	116.38
28	110.60	116.38
29	128.00	116.38
30	112.00	116.38
31	116.80	116.38
32	124.60	116.38
33	108.40	116.38
34	125.25	116.38
35	118.20	116.38
36	129.00	116.38
37	126.00	116.38
38	124.60	116.38
39	117.00	116.38
40	115.20	116.38
41	114.40	116.38
42	110.20	116.38
43	124.80	116.38
44	116.60	116.38
45	122.80	116.38
46	127.00	116.38
47	128.20	116.38
48	111.50	116.38
49	112.60	116.38


140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
1 2 3 4 5 6 7 8 9 10111213141516171819202122232425262728293031323334353637383940414243444546474849
Average Actual
Overall Average

Q2.B

Month-

If the individual monthly averages vary significantly from the overall average, we can say that the month has a significant impact. If most of the months vary significantly, we can conclude that month as an element affects daily case volume.

For January,

Ho: Average cases in January = yearly average

H1: Average cases in January != yearly average

Z-statistic test

Z= X-mean/ (pop std dev/sqrt(n))

Level of significance = 95%, two tailed test

Critical value is 1.96 and -1.96

If January's average is within critical region, do not reject Ho

Mean (yearly average) = 116.38

Month Nbr	Average Actual(X)	Std dev	N	X value when Z= 1.96	X value when Z= -1.96	Within critical region
1	114.77	11.69	22	121.27	111.50	Yes
2	114.33	9.27	21	120.34	112.42	Yes
3	116.50	13.18	22	121.89	110.87	Yes
4	118.57	12.89	21	121.89	110.87	Yes
5	117.59	13.21	22	121.90	110.86	Yes
6	124.10	11.58	21	121.33	111.43	No
7	114.48	6.89	21	119.33	113.43	Yes
8	124.35	8.00	23	119.65	113.11	No
9	112.11	8.16	9	121.71	111.05	Yes
10	113.69	10.04	16	121.30	111.46	Yes
11	118.19	25.56	21	127.31	105.45	Yes
12	104.64	37.52	22	132.06	100.70	Yes

For 10 of the 12 months, their average does not vary significantly from the yearly average. So, month as a factor does not have a significant impact on surgical volume. However, for June and August there is noticeable increase in average volume. To account this in the final model, we can add two dependent binary variables -is_june and is_august, it will be marked 1 for the given month and 0 for other cases.

Day of week -

Following the same approach for day of week

Level of significance = 95%,two tailed test

Critical value is 1.96 and -1.96

DOW	Average Actual(X)	Std dev	N	X value when Z= 1.96	X value when Z= -1.96	Within critical region
Mon	116.26	18.26	47	121.60	111.16	Yes
Tue	119.08	10.75	49	119.39	113.37	Yes
Wed	117.04	11.12	48	119.53	113.24	Yes
Thu	124.08	10.27	48	119.29	113.48	No
Fri	105.61	26.09	49	123.69	109.08	No

As two of the five DOW's have a significant impact on average case volumes. DOW as a factor has a significant impact on case volumes.

Week -

Following the same approach for week number

Level of significance = 95%,two tailed test

Critical value is 1.96 and -1.96

WeekNbr	Average Actual(X)	Std dev	N	X value when Z= 1.96	X value when Z= -1.96	Within critical region
1	114.60	7.99	5	123.39	109.38	Yes
2	115.00	9.65	5	124.84	107.92	Yes
3	113.20	12.14	5	127.02	105.74	Yes
4	113.00	9.40	5	124.62	108.14	Yes
5	125.20	11.62	5	126.56	106.20	Yes
6	130.20	2.79	5	118.82	113.94	No
7	92.00	45.97	4	161.44	71.33	Yes
8	121.40	10.65	5	125.72	107.05	Yes
9	121.40	8.59	5	123.91	108.85	Yes
10	127.80	5.04	5	120.80	111.97	No
11	94.40	40.49	5	151.87	80.89	Yes
12	70.80	48.09	5	158.54	74.23	No
13	112.00	13.46	5	128.18	104.58	Yes
14	106.20	8.84	5	124.13	108.63	No
15	114.20	10.85	5	125.89	106.87	Yes
16	124.40	5.39	5	121.11	111.66	No
17	115.20	8.38	5	123.72	109.04	Yes
18	106.60	7.81	5	123.23	109.53	No
19	113.80	8.82	5	124.11	108.65	Yes
20	123.00	4.94	5	120.71	112.05	No
21	120.60	6.31	5	121.91	110.85	Yes
22	116.00	14.52	5	129.11	103.66	Yes
23	116.40	7.31	5	122.79	109.97	Yes
24	104.80	8.54	5	123.87	108.89	No
25	124.60	13.41	5	128.14	104.63	Yes
26	117.20	11.96	5	126.86	105.90	Yes
27	121.00	14.79	5	129.35	103.42	Yes
28	110.60	8.33	5	123.69	109.08	Yes
29	128.00	8.65	5	123.96	108.80	No
30	112.00	12.13	5	127.02	105.75	Yes
31	116.80	10.24	5	125.36	107.40	Yes
32	124.60	15.34	5	129.83	102.93	Yes
33	108.40	7.31	5	122.79	109.97	No
34	125.25	7.15	4	123.39	109.37	No
35	118.20	8.38	5	123.72	109.04	Yes
36	129.00	11.47	5	126.44	106.33	No
37	126.00	14.52	5	129.11	103.66	Yes
38	124.60	8.82	5	124.12	108.65	No
39	117.00	9.97	4	126.16	106.61	Yes
40	115.20	4.62	5	120.43	112.33	Yes
41	114.40	4.32	5	120.17	112.60	Yes
42	110.20	6.85	5	122.39	110.38	No
43	124.80	8.59	5	123.91	108.85	No
44	116.60	4.03	5	119.91	112.85	Yes
45	122.80	6.46	5	122.05	110.72	No
46	127.00	8.02	5	123.42	109.35	No
47	128.20	5.49	5	121.20	111.57	No
48	111.50	9.50	4	125.69	107.07	Yes
49	112.60	6.86	5	122.39	110.37	Yes

31 of the 49 weeks don't vary significantly from the yearly average. WeekNbr is not a significant factor.

Q.3.A

The correlation matrix is

	T - 28	*T - 21*	*T - 14*	*T - 13*	*T - 12*	*T - 11*	*T - 10*	*T - 9*	*T - 8*	*T - 7*	*T - 6*	*T - 5*	*T - 4*	*T - 3*	*T - 2*	*T - 1*	*Actual*
T - 28	1																
T - 21	0.8947	1															
T - 14	0.766981	0.871427	1														
T - 13	0.761258	0.862506	0.975593	1													
T - 12	0.764272	0.84912	0.940374	0.977337	1												
T - 11	0.76968	0.839669	0.918844	0.955026	0.986618	1											
T - 10	0.744281	0.821875	0.91342	0.941554	0.962074	0.979289	1										
T - 9	0.718607	0.807351	0.924774	0.940412	0.941533	0.947764	0.973322	1									
T - 8	0.697891	0.794639	0.919929	0.931122	0.922158	0.918142	0.935192	0.971532	1								
T - 7	0.669865	0.769279	0.900452	0.91445	0.904064	0.89647	0.912204	0.955061	0.984829	1							
T - 6	0.669421	0.771311	0.890108	0.911955	0.912807	0.906488	0.918598	0.945678	0.969236	0.984542	1						
T - 5	0.679711	0.766765	0.863536	0.895554	0.919413	0.920257	0.922247	0.933364	0.948335	0.96	0.983981	1					
T - 4	0.685468	0.76623	0.846024	0.878267	0.910958	0.923938	0.927982	0.925826	0.930065	0.938392	0.963228	0.984911	1				
T - 3	0.686128	0.763745	0.845696	0.870565	0.893899	0.908863	0.926197	0.924528	0.92035	0.925503	0.946564	0.964317	0.984158	1			
T - 2	0.655022	0.742956	0.848112	0.862705	0.876955	0.885674	0.907966	0.922874	0.927708	0.934284	0.950649	0.959692	0.968785	0.983117	1		
T - 1	0.629432	0.718364	0.821478	0.83504	0.847387	0.851878	0.8712	0.895139	0.909233	0.918124	0.927954	0.937331	0.943132	0.950928	0.970063	1	
Actual	0.60829	0.702459	0.800877	0.81273	0.818714	0.819855	0.842193	0.87289	0.887675	0.895779	0.8989	0.902827	0.90604	0.913242	0.93643	0.964727	1

Correlation coefficient(r) is the measure of strength of linear association between two variables.

Actual and T-1 are highly correlated and their r is the highest (0.964727) of all other Actual and T-x combinations. T-1 is most useful in predicting case volume.

Q.3.B One of the main assumptions of a linear model is that independent variables should not be correlated. Any two of the T-x variables cannot be used as predictors as they are highly correlated. The least correlation for any combination of two T-x variable is 0.629 and their strength of association is large. It is not useful to include multiple booking dates in the prediction model.

Q.4.

1.Simple Linear Regression

To start with, we built a series of linear regression models with T-x variable as single predictor and Actual as the outcome variable. For T-1 as predictor, the R Square is 0.93(i.e., this explains 93% of the variance in Actual). The R square goes on decreasing as we move away from T-1 till T-28. For T-28, it is 0.37. Simple Linear regression with one predictor provides accurate results. Lets move forward to see if multiple regression adds any benefits.

T-1 Simple linear regression

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.964726859							
R Square	0.930697912							
Adjusted R Square	0.930407945							
Standard Error	4.650686599							
Observations	241							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	69421.57595	69421.57595	3209.669535	1.5737E-140			
Residual	239	5169.303716	21.62888584					
Total	240	74590.87967						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.18269892	1.880881367	5.9454568	9.70992E-09	7.477476593	14.88792124	7.477476593	14.88792124
T - 1	0.956282799	0.016879368	56.65394545	1.5737E-140	0.923031466	0.989534131	0.923031466	0.989534131

2.Multiple Linear Regression

2.A. With two or more T-x variables

We noticed that there is multicollinearity among T-x variables. This breaks an assumption of linear model. However, there is a workaround using VIF(Variance Inflation Factor) values.

Firstly, using stepwise regression, we get the best combination of T-x variables. The stepwise regression iteratively adds and removes predictors and finds the combination of predictor variables giving the best performing model (lowest prediction error).

```
Final Model:
Vandert$Actual ~ Vandert$`T - 7` + Vandert$`T - 6` + Vandert$`T - 1`
```

```
                  Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                          224   4935.083 761.6580
2   - Vandert$`T - 8`  1  0.1040491       225   4935.187 759.6631
3  - Vandert$`T - 10`  1  0.2559364       226   4935.443 757.6756
4  - Vandert$`T - 12`  1  1.0051187       227   4936.448 755.7247
5   - Vandert$`T - 5`  1  1.2434674       228   4937.692 753.7854
6   - Vandert$`T - 4`  1  3.3293546       229   4941.021 751.9478
7  - Vandert$`T - 28`  1  7.0817260       230   4948.103 750.2930
8  - Vandert$`T - 13`  1 15.6274538       231   4963.730 749.0529
9  - Vandert$`T - 14`  1  4.0987942       232   4967.829 747.2519
10  - Vandert$`T - 2`  1 14.7995991       233   4982.629 745.9687
11  - Vandert$`T - 3`  1 11.5208562       234   4994.150 744.5253
12 - Vandert$`T - 21`  1 20.3547763       235   5014.504 743.5056
13  - Vandert$`T - 9`  1 29.0680455       236   5043.572 742.8986
14 - Vandert$`T - 11`  1 14.3003110       237   5057.873 741.5809
```

Secondly, we check for multicollinearity among selected predictors and remove the one with largest VIF (>10). VIF is the variance in the model with predictor variables combined divided by the variance with single predictor. The basic rule of VIF is that it is acceptable if it less than 10.

```
Vandert$`T - 7` Vandert$`T - 6` Vandert$`T - 1`
      32.753782       37.033082        7.233965
```

Interpreting from above results, we remove the predictor T-6 as it has the largest VIF and recalculate.

We check for multicollinearity for the combination of T-1 and T-7

```
Vandert$`T - 7` Vandert$`T - 1`
       6.367458        6.367458
```

Here we get the acceptable value of VIF (<10). This model meets the assumption of multicollinearity. Finally, we build the model and results are:

```
Call:
lm(formula = Vandert$Actual ~ Vandert$`T - 7` + Vandert$`T - 1`,
    data = Vandert)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.4954  -2.9540  -0.0049   2.7722  16.9760 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     11.55885    1.89295   6.106 4.12e-09 ***
Vandert$`T - 7`  0.07005    0.04696   1.492    0.137    
Vandert$`T - 1`  0.89810    0.04248  21.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.639 on 238 degrees of freedom
Multiple R-squared:  0.9313,	Adjusted R-squared:  0.9308 
F-statistic:  1614 on 2 and 238 DF,  p-value: < 2.2e-16
```

Based on the P-value, we notice T-7 does not have a significant impact. So, we continue with Simple Linear Regression with single predictor.

2.B. With one T-x variable and another categorical variable (DOW,Month)

Another approach is to add in categorical variables to existing T-x single linear regression model. In this case, it is unnecessary to add such factors as they don't capture any additional effect/trend that the T-x variables miss out.

```
Call:
lm(formula = Vanderfull$Actual ~ Vanderfull$DOW + Vanderfull$`T - 1`,
    data = Vanderfull)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1881  -2.9753  -0.1359   2.7254  16.3623

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.14475    1.88106   5.925 1.11e-08 ***
Vanderfull$DOWMon  -2.30959    0.96930  -2.383  0.01798 *
Vanderfull$DOWThu  -1.07112    1.00236  -1.069  0.28635
Vanderfull$DOWTue  -2.61819    0.97616  -2.682  0.00783 **
Vanderfull$DOWWed  -1.16394    0.96280  -1.209  0.22791
Vanderfull$`T - 1`  0.96961    0.01809  53.594  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.598 on 235 degrees of freedom
Multiple R-squared:  0.9334,    Adjusted R-squared:  0.932
F-statistic: 658.7 on 5 and 235 DF,  p-value: < 2.2e-16
```

For example, DOW is added as a predictor. It provides the information that Thursday's must be higher than average DOW and Friday's lesser than average DOW. However, the T-x variable used in regression easily captures this pattern by itself. The T-x variables are higher than average T-x for Thursday's and lesser than average T-x for Friday's. Adding DOW will lead to overfitting and it is not advisable. The R square value with DOW and T-1 is 0.9334 which is slightly higher compared to taking T-1 alone which is 0.9307.

3.Choosing the best Linear Model

In an ideal case, the hospital would expect a highly accurate model as early as possible for scheduling. However, in reality, there is a trade off between accuracy and time of prediction. The higher the accuracy the closer T-x is to the actual date.

T-x (X=)	RMSE
1	4.63
2	6.17
3	7.16
7	7.81
10	9.48
14	10.53
28	13.96

RMSE (Root Mean Squared error) is the squared root of mean of the square of residuals. RMSE is a measure of accuracy, to compare prediction errors of different models.

As we expected, the model accuracy steadily increases as we get closer to actual date. Depending on the scheduling time required and the penalty costs incurred for inaccurate prediction, the hospital will have to decide the best T-x model.