Executive summary

In order to give correct advice to Mr. Roddey, various statistical models were made to verify the influencing factors. Further a prediction model was also made.

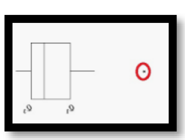
As per my analysis, when Nobel had pitched, the sales had increased. There is no doubt about that but Nobel played strategically against those teams and on those days [Saturday and Sunday/times [night games] when the sale was due to increase. For example he played **9** out of **16 games at night.** And he played **60%** of games on weekends. But whether these factors affect sales or no, will be analyzed in my model.

Nobel	Weeday	Weekend	Grand Total	Percentage payed on weekends
Without Nobel	39	20	59	33%
With Nobel	10	6	16	60%
Grand Total	**49**	**26**	**75**	

Further in my model, I will also validate the feelings of Mr. Nobel as encircled. As per Nobel, following was said in a magazine:

Nobel also argued that he had the ability to attract people to the ball park. He had been quoted in *Sports Illustrated*[5] as saying: "I'm not saying anything against Rick Langford or Matt Keough [fellow A's pitchers]…but I filled the Coliseum last year against Tommy John [star pitcher for the Yankees]." The implication was that Nobel felt he did indeed personally attract people to the games.

As per my analysis, the **"feeling"** was incorrect that Nobel personally attracted rather he might have conceivably attracted crowd. Because he has played almost **90% (13 out of 16)** which were not based on promotion. Hence there is a doubt.



Lastly Nobel's salary cannot be based on the average, because there is an outlier (team Yankees) which if removed will level the sales – with or without Nobel. That shows the coliseum was not filled because of him but because of the game with popular opposition.

Thus in my analysis, I will validate the association of all Factors with total sales and will provide statistical evidence whether the factor is significant or no.

Detailed Summary.

Data Extraction

After reading the business problem, following basic descriptive analysis and data extraction was performed

1. Date & number was removed as redundant column.
2. Since divisions were mentioned in the exhibit, data was divided into divisions. Thus we had new column WestDiv_EastDiv. Thus it was identified that White Sox was one team without **any rank**.
3. Factors affecting attendance was analyzed and following new features were created:
    a. **Weekday_Weekend**: Weekday as 0 and Weekend as 1. This was done using vlookup.
    b. **Double header** was included as new column
    c. OppTeam: Two new features were created
        i. OppTeam_rank
        ii. **Oppteam_*team name*** *(this was done as part of encoding through get_dummies)*

Apart from above, I created Pivot table to see the data and verified the average reported by the manager with or without Nobel and then realized that I should create one more column called: Increase from previous play. We will see the usage later on.

```
['pos', 'gb', 'pos_gb', 'temp', 'prec', 'oppteamrank', 'doubleheader_1',
'oppteam_Boston', 'oppteam_California', 'oppteam_Cleveland',
'oppteam_Detroit', 'oppteam_Kansas City', 'oppteam_Milwaukee',
'oppteam_Minnesota', 'oppteam_Seattle', 'oppteam_Texas',
'oppteam_Toronto', 'oppteam_White Sox', 'oppteam_Yankees', 'tog_2',
'tv_1', 'promo_1', 'nobel_1', 'wdzero_weone_1', 'westzero_eastone_1']
```

Head On → Initial Descriptive and Statistical Analysis

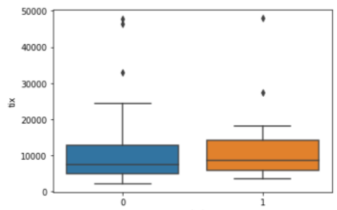
I took the problem of identifying Nobel's influence by adopting these three statistics

1. Outlier analysis: Python
2. Two sample t test: Python
3. Regression: Xls
4. Correlated various values extracted values.
5. Did two sample test or Annova for almost each of the factors.

Outlier analysis: Python

```
In [408]: sn.boxplot(x = 'nobel', y = 'tix', data = raw_df)
          #sn.barplot(x='dow',y='tix', hue = 'nobel',data = raw_df)

Out[408]: <matplotlib.axes._subplots.AxesSubplot at 0x20cad165eb8>
```



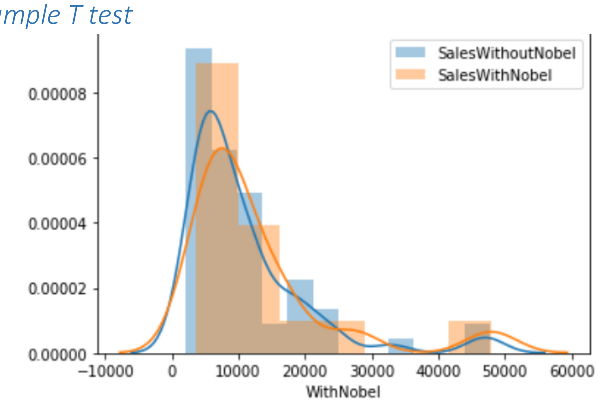
As mentioned in the summary, there are situations when the sales are high and that outlier is also influencer, whenever the match is played with Yankees, the sales are high. But this is just outlier analysis and is a claim. We will prove the claim statistically.

Two Sample T test: Python

Created Distribution Plot

```
influence_of_nobel_on_sales = pd.read_csv("InfluenceOfNobelOnSales.csv")
with_n = pd.DataFrame()
without_n = pd.DataFrame()
with_n = influence_of_nobel_on_sales.WithNobel
with_n_final = with_n.dropna()
#with_n_final
without_n = influence_of_nobel_on_sales.WithoutNobel
without_n_final = without_n.dropna()
#influence_of_nobel_on_sales
sn.distplot(without_n_final,label = 'SalesWithoutNobel')
sn.distplot(with_n_final,label = 'SalesWithNobel')
plt.legend();
```

Two Sample T test



**We can observe that the distribution is overlapping and hence the impact of Nobel starting as a pitcher is not contributing towards ticket sales. This can be verified in two ways: *

> 1. Annova
> 2. Two sample t test.

Since the samples are only two, t test and annova is giving p value of .514 which means that sales with and without Nobel is not significantly different.

```
[190]: scistats.ttest_ind(with_n_final,without_n_final)

[190]: Ttest_indResult(statistic=0.6553210611376503, pvalue=0.5143209387067114)
```

Correlation: xls

In order to identify the influencing factor, regression will be done but before we perform regression, we should see the correlation.

	TIX	DoubleHeader	OPP	POS	GB	DOW	TEMP	PREC	TOG	TV	PROMO	NOBEL	OppTeamRank	WDzero_Weone	Westzero_Eastone
TIX	1														
DoubleHeader	0.1719	1.0000													
OPP	-0.1122	-0.2493	1.0000												
POS	-0.1149	-0.0781	-0.2059	1.0000											
GB	0.0748	-0.0698	0.1848	-0.1521	1.0000										
DOW	-0.0070	0.1096	-0.0564	-0.1044	-0.1230	1.0000									
TEMP	-0.0608	-0.0917	-0.1061	0.0524	0.6571	-0.1352	1.0000								
PREC	-0.0974	0.1906	-0.0029	-0.1816	-0.1613	0.1757	-0.2901	1.0000							
TOG	0.1287	0.1102	-0.1116	0.1257	0.0914	-0.5564	0.1135	-0.0599	1.0000						
TV	-0.0979	0.0423	0.1241	-0.0825	-0.1692	0.1974	-0.0775	0.1340	-0.2727	1.0000					
PROMO	0.2666	-0.1350	-0.0249	-0.0025	0.1176	-0.0021	0.1670	-0.0935	-0.0169	-0.0607	1.0000				
NOBEL	0.0765	0.2063	-0.0928	0.0329	0.0045	0.0821	-0.0338	0.0598	0.0860	0.0080	0.0195	1.0000			
OppTeamRank	-0.2765	0.2149	-0.2994	-0.0145	-0.3236	0.0183	-0.2902	0.2402	-0.0053	0.0853	-0.0187	0.1851	1		
WDzero_Weone	0.0462	-0.0083	-0.0102	-0.1090	-0.1384	0.8050	-0.1077	0.1373	-0.6999	0.3345	0.1105	0.0310	0.020420767	1	
Westzero_Eastone	0.3183	-0.0118	0.1466	-0.3458	0.0503	-0.0177	-0.0569	0.0599	0.0150	-0.0558	0.2284	0.0443	0.124846147	0.02691747	1

Weekend and time of game are related but TOG impacts more towards TIX.

Lower rank team, crowd decreases.

More promotion, will lead more sales

This is very important relation but it is related with oppteam rank

Feature – Annova & Inferences

In order to see the impact, Annona was performed for each variable.

Weekend (Saturday/Sunday) - 0 Weekday – 1

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0	49	535026	10919	1E+08
1	26	308293	11857	8E+07

ANOVA

Source of Vari	SS	df	MS	F	P-value	F crit
Betwee	1E+07	1	1E+07	0.1562	0.69379018	3.972038
Within (	7E+09	73	1E+08			

Though p value is less but their averages are good for weekends – which is day 6 & day 7 combined.

Time of day

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
TOG_day	39	391902	10049	7E+07
TOG_night	36	451417	12539	1E+08

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1E+08	1	1E+08	1.2304	0.27096947	3.972038
Within Groups	7E+09	73	9E+07			
Total	7E+09	74				

Opposition Team (oppteam)

SUMMARY

Groups	Count	Sum	Average	Variance
Baltimore	6	70518	11753	44127767
Boston	6	78140	13023.33	15325115
California	5	39856	7971.2	5480014
Cleveland	6	64669	10778.17	52718197
Detroit	5	47698	9539.6	24205264
Kansas Cit	6	69913	11652.17	49512424
Milwauke	6	49763	8293.833	26206014
Minnesota	7	70316	10045.14	57075863
Seattle	6	42792	7132	25748591
Texas	6	35523	5920.5	11467059
Toronto	5	40766	8153.2	15905651
White Sox	6	31140	5190	5191512
Yankees	5	202225	40445	93357355

ANOVA

ce of Varia	SS	df	MS	F	P-value	F crit
Between (	4.96E+09	12	4.13E+08	12.49203	2.03E-12	1.911926
Within Gr	2.05E+09	62	33060189			
Total	7.01E+09	74				

Sales also depends against whom the A's are playing. In business case it is written that Yankees were favorite and it can be seen here with the count and with the p value of the group.

Doubleheader

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0	69	742055	10754	8E+07
1	6	101264	16877	2E+08

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2E+08	1	2E+08	2.2221	0.14036018	3.972038
Within Groups	7E+09	73	9E+07			
Total	7E+09	74				

Position

SUMMARY

Groups	Count	Sum	Average	Variance
1	9	97305	10812	42709716.5
2	25	3E+05	11677	100531296
3	23	3E+05	13367	152777393
4	9	65739	7304.3	33264017.8
5	6	67710	11285	45132966
6	1	2140	2140	#DIV/0!
7	2	11043	5521.5	136764.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4E+08	6	7E+07	0.68292729	0.6639311	2.23521
Within Groups	7E+09	68	1E+08			
Total	7E+09	74				

If you look at top 3 rows: I believe that position of A also matters in collecting sales. The better they perform, the more they attract crowd.

Promotion

Anova: Single Factor

SUMMARY

Groups	Count	Sum	\verage	⁄ariance
0	62	6E+05	10065	8E+07
1	13	2E+05	16870	1E+08

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	497774506	1	5E+08	5.584	0.020794573	3.972
Within Groups	6507824062	73	9E+07			
Total	7005598568	74				

Promotion matters in increasing the sales – it has high correlation with tix and above stats prove the same.

GamesBehind

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0	9	97305	10811.7	4.3E+07
1	8	58594	7324.25	5.1E+07
2	5	40438	8087.6	6E+07
3	2	15037	7518.5	1.8E+07
6	2	28937	14468.5	4371925
7	6	123496	20582.7	2.6E+08
8	1	46294	46294	#DIV/0!
9	1	17666	17666	#DIV/0!
10	1	4899	4899	#DIV/0!
11	4	32450	8112.5	3504569
12	14	124303	8878.79	1.5E+07
13	4	31339	7834.75	1.3E+07
14	3	12115	**4038.33**	951344
15	5	45051	9010.2	1.4E+07
16	2	21831	10915.5	227813
17	6	131390	**21898.3**	2.6E+08
18	1	2443	2443	#DIV/0!
19	1	9731	9731	#DIV/0!

ANOVA

ce of Varia	SS	df	MS	F	P-value	F crit
Between (	3.2E+09	17	1.9E+08	2.75841	0.00218	1.80518
Within Gr	3.8E+09	57	6.7E+07			
Total	7E+09	74				

Temperature

MMARY				
roups	Count	Sum	Average	Variance
55	2	9990	4995	8712
56	1	4141	4141	#DIV/0!
57	5	60453	12090.6	48435799
58	4	51157	12789.25	73242798
59	5	36215	7243	25270374
60	8	108353	13544.13	2.41E+08
61	5	57770	11554	40021327
62	10	157248	15724.8	2.05E+08
63	11	134267	12206.09	50903817
64	7	79297	11328.14	2.44E+08
65	9	90186	10020.67	13169011
66	2	17066	8533	15724832
67	2	16585	8292.5	26521045
69	3	17522	5840.667	5700337
70	1	3069	3069	#DIV/0!

OVA						
of Varia	SS	df	MS	F	P-value	F crit
tween (	6.76E+08	14	48253494	0.457376	0.946514	1.8602

Opposition Team Rank

F17 | 0.0000686319329799658

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	0	6	31140	5190	5191512		
6	1	11	272138	24739.82	2.88E+08		
7	2	6	70518	11753	44127767		
8	3	13	120079	9236.846	40282858		
9	4	12	113663	9471.917	25937418		
10	5	5	47698	9539.6	24205264		
11	6	11	104525	9502.273	30699939		
12	7	11	83558	7596.182	19520969		
13							
14							
15	ANOVA						
16	Source of Variation	SS	df	MS	F	P-value	F crit
17	Between Groups	2.51E+09	7	3.58E+08	5.341606	6.86E-05	2.149653
18	Within Groups	4.5E+09	67	67109065			
19							
20	Total	7.01E+09	74				

Rank 0! What is this: This is whitesox which is without any division.

Did Nobel play against this team? – Yes he played

	Column Labels					
	Average of TIX		Count of NOBEL		Total Average of TIX	Total Count of NOBEL
Promotion	0	1	0	1		
Without Nobel	4531.6		5		4531.6	5
With Nobel		8482		1	8482	1
Grand Total	4531.6	8482	5	1	5190	6

Did he attract the crowd? May be not –Since it was night game with promotion.

	Column Labels					
	Average of TIX		Count of NOBEL		Total Average of TIX	Total Count of NOBEL
Promotion	No promo	Promo	No promo	Promo		
Without Nobel	4531.6		5		4531.6	5
Day	5251		3		5251	3
Night	3452.5		2		3452.5	2
With Nobel		8482		1	8482	1
Night		8482		1	8482	1
Grand Total	4531.6	8482	5	1	5190	6

West – East Division

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
EastDiv	39	553779	14199.46	1.39E+08
UknownDiv	6	31140	5190	5191512
WestDiv	30	258400	8613.333	31948589

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	7.68E+08	2	3.84E+08	4.433597	0.015281	3.123907
Within Groups	6.24E+09	72	86630910			
Total	7.01E+09	74				

Though P value is low enough but we can't consider this feature as we don't know division of Whitesox.

Multiple Linear Regression Prediction Model

Using the data extracted above (section: data extraction), following steps were taken:

1. Created regression with all variables
2. Analyzed VIF
3. Based on VIF, Correlation; new features were selected.
4. Created 2 new models
5. Checked diagnostics and RMSE & R2 score

Model Summary

Created two models to explain the significance of features which are associated with TIX. Also to explain the influence:

Model1_Yank_WEND_TOG

1. Yankees are associated with Sales. As per the business case, Yankees was famous and Roddey believed that it would attract large crowd. His belief is correct and that is seen in model. Also Nobel is incorrect when he says that he attracted large crowd when he was playing against Yankees pitcher. Nobel has made this statement in magazine and hence - Roddey's belief is correct and Nobel's belief is incorrect

2. Roddey's belief that Weekends attract more crowd is also significant.

3. Thirdly time of game impacts the sale. This is also seen statistically.

Model2_promo_tog

Since Yankees is representing all leverage values, we will need another model to prove the significance of other variables such as:

1. Promotion – Annova single factor significant / .27 correlation with TIX

2. Position (which is performance of A) / .11 correlation with TIX

3. Rank of opposition team - Annova single factor significant / .3 correlation with TIX

4. Games behind the opposition team.- Annova single factor significant / not correlated

Model Detailed Calculation –

Model1: Tix (Y) = Yankees, Weekday_Weekend, Time of game

Tog_2 means night game: If all are constant and match is played on weekend then increase is around 9.5k

Wdzero_weone_1 means this is weekend: If all are constant and match is played on weekend then increase is around 8k

Boston: This is almost **insignificant** because if removed, then also the model R square, plots to do not change.

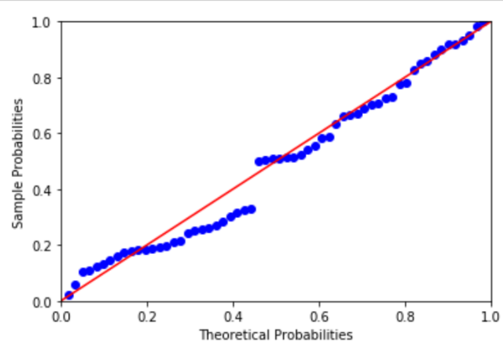
And of course Yankees is highly significant

Model:	OLS	Adj. R-squared (uncentered):	0.826
Dependent Variable:	tix	AIC:	1218.7109
Date:	2020-06-02 17:49	BIC:	1227.0883
No. Observations:	60	Log-Likelihood:	-605.36
Df Model:	4	F-statistic:	72.19
Df Residuals:	56	Prob (F-statistic):	1.94e-21
R-squared (uncentered):	0.838	Scale:	3.6384e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
oppteam_Yankees	29391.5914	3156.9319	9.3102	0.0000	23067.4968	35715.6860
oppteam_Boston	5314.9587	2585.1413	2.0560	0.0445	136.2984	10493.6190
wdzero_weone_1	8818.7042	1281.2618	6.8828	0.0000	6252.0283	11385.3801
tog_2	9537.6131	1249.0526	7.6359	0.0000	7035.4600	12039.7662

PP Plot

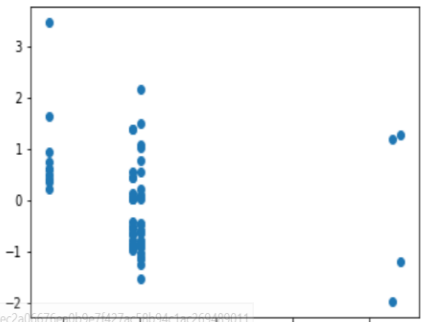
```
sm.ProbPlot(ipl_model_1_2.resid).ppplot(line = '45')
```



Residual

```
#sm.ProbPlot(ipl_model_l_2.resid).ppplot(line = '45')
plt.scatter(get_standarized_values(ipl_model_l_2.fittedvalues),get_standarized_values(ipl_model_l_2.re
```

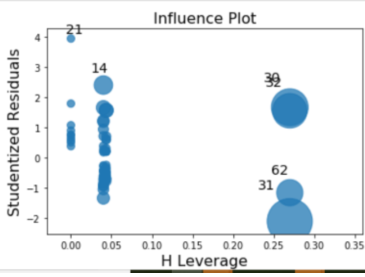
Out[233]: <matplotlib.collections.PathCollection at 0x27b8a153dd8>



Leverage

In [234]:
```
#Levereage cut off : 1.35
from statsmodels.graphics.regressionplots import influence_plot
influence_plot(ipl_model_l_2,'true')
```

Out[234]:



All are Yankees with leverage cut off

29]:

	number	date	doubleheader	tix	increase	opp	pos	gb	pos_gb	dow	temp	prec	tog	tv	promo	nobel	oppteam
30	31	13-Jun	1	47768	42140	4	3	7	21	5	60	0	2	0	0	0	Yankees
31	32	14-Jun	0	27312	-20456	4	3	7	21	6	63	0	1	0	0	1	Yankees
32	33	15-Jun	0	46294	18982	4	3	8	24	7	64	0	1	0	1	0	Yankees
62	63	26-Aug	0	32905	-15041	4	2	17	34	2	62	0	2	0	1	0	Yankees

Predicted R2_score

Rsquare is .8 but on test it is .4: thus model is over fitting but not highly over-fit as value (.4) is less. Secondly aim is to make inference rather than prediction.

```
[230]: #predict y for each row of test data set. But we will predict using those x which were used to create 
       pred_y = ipl_model_1_2.predict(test_x[train_x_final_1.columns])

       actual_y_pred_y=pd.DataFrame()
       actual_y_pred_y['pred_y'] = pred_y
       actual_y_pred_y['actual_y'] = test_y
       #actual_y_pred_y
       from sklearn import metrics
       #import numpy as np
       #RMSE - smaller the better
       #np.sqrt(metrics.mean_squared_error(pred_y,test_y))
       #R suare - compare with model's R square.
       #If less, then model is overfitting. If high, then model is underfitting
       np.round(metrics.r2_score(pred_y,test_y),2)

[230]: 0.4
```

Model2– Tix (Y) = Promotion, TOG

There two Sub models created:

- GB, OppTeamRank, Promotion – Later on we discard this model
- Promotion & TOG

Since Yankees has high leverage value, trying to remove Yankees from train set and creating a model:

Model without Yankees

Model:	OLS	Adj. R-squared (uncentered):	0.701
Dependent Variable:	tix	AIC:	1132.3133
Date:	2020-06-02 20:58	BIC:	1138.3894
No. Observations:	56	Log-Likelihood:	-563.16
Df Model:	3	F-statistic:	44.74
Df Residuals:	53	Prob (F-statistic):	1.50e-14
R-squared (uncentered):	0.717	Scale:	3.3596e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
gb	392.4514	98.8944	3.9684	0.0002	194.0943	590.8084
oppteamrank	958.5033	222.9309	4.2996	0.0001	511.3605	1405.6461
promo_1	5644.0576	2041.9415	2.7641	0.0078	1548.4417	9739.6735

Though R squared is .7 and the residual plot also looks ok, but this is not giving correct inference. As team rank increases, the sale also increase. That is not true as seen in data and explained in business problem because people want to see winning teams. Similarly, stats are not true for GB. This inappropriateness is seen in **<u>negative r2_score of "– 20" and in cone shaped residual plot.</u>**

Thus, we will **discard this model and create another one based on** promotions and tog (tog is already **considered significant in Model1**).

[295]:

Model:	OLS	Adj. R-squared (uncentered):	0.523
Dependent Variable:	tix	AIC:	1157.4765
Date:	2020-06-02 21:06	BIC:	1161.5272
No. Observations:	56	Log-Likelihood:	-576.74
Df Model:	2	F-statistic:	31.72
Df Residuals:	54	Prob (F-statistic):	7.75e-10
R-squared (uncentered):	0.540	Scale:	5.3558e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
tog_2	8396.9018	1546.2868	5.4304	0.0000	5296.7833	11497.0202
promo_1	10126.8393	2395.4973	4.2274	0.0001	5324.1564	14929.5222

Omnibus:	2.944	Durbin-Watson:	1.372
Prob(Omnibus):	0.229	Jarque-Bera (JB):	2.023
Skew:	0.405	Prob(JB):	0.364

For this p-p plot is normal and RMSE is 0. Also the model gives better inference. Night games and promotions are influencing sales. Though model is over fit, yet inferences are in line with Annova test done earlier.

Case Questions

Does Mark Nobel increase attendance? If so, how much is this increase worth to the Oakland A? (20 points)

As identified in previous models, following features were associated with the sales and those are:

1. Opposition Team (that too Yankees had huge influence) as seen in Annova and in regression.
2. Time of game & Week ends. – This is also seen in Model 1.

Since Model 1 was based on Yankees which was having high impact, we removed Yankees to see the impact of other features and identified following:

1. Promotion: This had significance in two sample t test and in regression too.

If we remove above influencing features (PROMO (1), OppTeam(Yankees) and TOG (night) from data, then we will see this calculation:

PROMO	0		
OppTeam	(Multiple Items)		
TOG	1		
Row Labels	Average of TIX	Count of NOBEL	Sum of TIX
0	7807.923077	26	203006
1	9088.4	5	45442
Grand Total	**8014.451613**	**31**	**248448**
1280.476923	Average difference		
4609.716923	Multiplied by 3.6		
73755.47077	Worth of Nobel		

Thus worth of Nobel is determined by removing the factors which influence the TIX. Now one could argue that Nobel has played more games over the weekend and hence his worth should not be defined by just these parameters (i.e. TOG, PROMO, OPP TEAM). This is true but we need to consider the performance of Nobel [i.e. HIGH ERA] and we need to consider the fact that Nobel attracts more crowd even without Promo.

9 out 16 appearance of Nobel are in games which have no promotions but still he is attracting crowd on an average when he plays.

And if we remove Yankees then also average crowd is more when he plays even without promotion. Please see below:

	Column Labels				
	No promotion		Promotion		To
Row Labels	Count of NOBEL	Average of TIX	Count of NOBEL	Average of TIX	
0	48	8353.958333	8	14093.125	
1	11	9089.818182	3	9123.666667	
Grand Total	**59**	**8491.152542**	**11**	**12737.81818**	

Hence, Nobel should be compensated properly because his worth is 73K approx. and he has ability to influence total sales of tickets.

If you used hypothesis testing or regression model to answer question 1, comment on the appropriateness of both these techniques. (5 points)

Regression will answer the association of Nobel with TIX in a joint way that means, along with other features. Hypothesis testing will answer the association of Nobel with TIX in simple one way.

Secondly, regression gives us an opportunity to validate the association of Nobel and other features by predicting the model. Regression gives us prediction model as a byproduct of the complete exercise.

Lastly, I could have changed the critical value (80% significance from 95%) and analyzed the association of Nobel with TIX. This is very much possible in regression and it can be done for all other features together with Nobel.

Thus regression gives more insights and flexibility.

What should be the ideal salary for Mark Nobel? (2 points)

Ideal salary cannot be calculated because in the problem statement nothing is given to calculate the.

I would calculate his salary based on proportions.

When worth was 105650 then his demand for salary 600000			
When his worth is 73755, propotionately, his new salary could be			
New Salary	418864.1742		
Approx new salary	418,000		

His salary should be 418K dollars instead of 600k dollars.

List all the insights based on your data analysis. (3 points)

Most of the analysis is already mentioned but lastly, I want to mention two things:

1. When Nobel started 22 out of 33 games were won. This means that in 1980, probably 10 out 16 games were won when he would have started.
2. In my data extraction, I have mentioned that I have extracted one feature – Increase. This is increase in sales with respect to previous game. There is a correlation between increase and Nobel but there is NO significance of Nobel associated with Increase. In fact none of the features were associated, via regression. But we can see that Nobel had resulted in higher average increase as seen in below chart:

When games were played on weekdays and were played at day times, the average increase as compared to previous day was more when Nobel had started the pitch. As seen below:

Average increase was 47 with respect to previous day even when there was no promotion

Row Labels	Count of NOBEL	Average of TIX	Average of Increase
⊟0	**12**	**6667.583333**	**4.505727466**
No promotion	11	6762.090909	0.795910845
Promotion	1	5628	45.3137103
⊟1	**1**	**3598**	**47.27793696**
No promotion	1	3598	47.27793696
Grand Total	**13**	**6431.461538**	**7.795897428**

Note: I have filtered out night games and weekends in my pivot table.

Lastly,

Even though I have calculated the worth but I have not considered illegalities in pitching as mention in case and nor measured genuine of Nobel's hard work, performance.

Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?
Dennis Prager

Thank you for reading through!