Name:	**Mansi Dobariya**
Enrollment Number:	**AU1841131**
Course:	**TOD206 - Industrial Statistics**
Major in:	**B.Tech(ICT)**
Topic:	**Assignment -3**

Ahmedabad University

a)Topic title
Cricket score prediction of player

b) Introduction (Problem statement, Overview, Description, Objectives, etc.)
Problem statement:

-Various brand companies offer discounts, prizes for marketing their products during the IPL months to the customers who correctly predict the winner.It Helps in making decision whether to go to the stadium or not on your favourite team's match.Many factors affect in prediction of points awarded to player,i.e. Number of wickets taken by players,fours,sixes, catches,winning toss, batting side, Home ground advantages, player wise performance etc. An early prediction is always helpful for team management to work on their plans quickly and improve team performance and enhance the chances of winning the game.With use of **multiple linear regression** to predict points of each player in the league and then the League decides which player in which team based on the overall strength of each team based on the past performance of the players who have appeared most for the team.

What is the predicted score of one player for 7 wickets taken by him , 4 fours ,6 sixes and 5 catches during IPL 2017?

Overview:

This report contains detailed statistical analysis of Multiple Regression with the help of dataSet.Including Data collecting,Cleaning,feature selection and testing the model.Data collected from the official website of IPL 2017-2018. Using significance information we can say that model or variable is significant or not. And with the help of a backward elimination method we conclude that this model is the best model to predict the score. We can reach the same conclusion whether using an equation or plot in each part mentioned below.

Description:

-There are various ways a player can be awarded points for their performance in the field. The official website of IPL has a Player Points section where every player is awarded points mainly based on these 4 features:

(i) number of wickets taken
(ii) number of fours
(iii) number of sixes
(iv) number of catches,

To find out how IPL management was assigning points to each player based on these 4 features, a multiple linear regression was used on the players' points data.

Objectives:

-For this problem with four independent variables(X_1 , X_2 , X_3 , X_4) and Dependent variable (**Y**) the multiple linear regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

where,

Y: points awarded to a player

X_1: no. of wicket taken by player

X_2: no. of four

X_3: no. of six

X_4: no. of catch

β_0: the population intercept

β_1: per wicket weight(slope coefficient)

β_2: per four weight(slope coefficient)

β_3: per six weight(slope coefficient)

β_4: per catch weight(slope coefficient)

-Virtually all regression analysis of business data involves sample data, not population data. As a result, β_0, β_1, β_2, β_3 and β_4 are unattainable and must be estimated by using the sample statistics, b_0, b_1, b_2, b_3 and b_4.

-Hence the equation of the regression line contains the sample y-intercept, b_0 and The sample slope b_1, b_2, b_3 and b_4.

Mathematical form:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

Where,

$\hat{Y}$: Estimated (or predicted) Y value

b_0: Estimate of the regression intercept

b_1: Estimate of the regression slope per wicket

b_2: Estimate of the regression slope per four

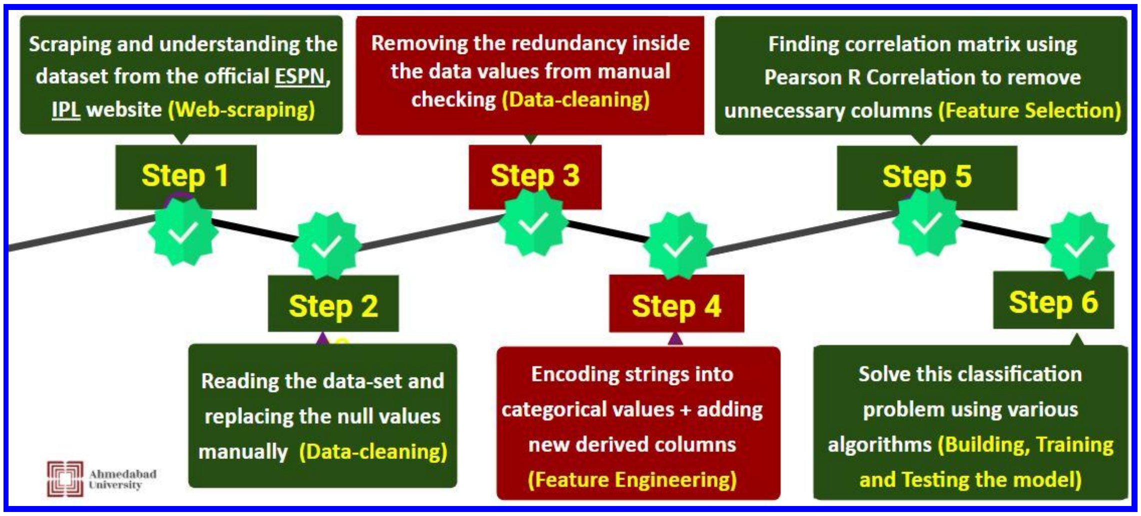
b_3: Estimate of the regression slope per six

b_4: Estimate of the regression slope per catch

To determine the equation of the regression line for a sample of data, we must determine the values for b_0, b_1, b_2, b_3 and b_4.

c) Statistical Data

-I choosed Scraping and understanding the dataset from the official ESPN, IPL website with my prior knowledge of previous courses and references. After that I cleaned the data and generated random valued columns and got 25 sample size data. Flow of process is given below:



Dataset in Excel

PLAYER NAME	Points awarded to player	No.of wicket taken	No.of Four	No.of Six	No.of catch
	Y	X1	X2	X3	X4
Saurabh Tiwary	33	0	8	3	1
Nathan Coulter-Nile	101.5	5	4	0	2
Mohammed Siraj	130.5	11	2	0	4
Chris Green	6	0	0	0	0
Virat Kohli	103.5	0	23	11	3
Dinesh Karthik	86.5	0	20	4	9
Kane Williamson	115	0	26	10	6
James Pattinson	126.5	11	2	0	2
Sandeep Warrier	8	0	0	0	0
Rahul Tripathi	95	0	21	10	3
Nicholas Pooran	162.5	0	23	25	7
Tushar Deshpande	50.5	3	2	1	1
Ajinkya Rahane	47	0	12	2	4
Murali Vijay	10	0	4	0	0
Josh Philippe	26	0	9	1	0
Marcus Stoinis	236.5	13	31	16	3
Kuldeep Yadav	25	1	1	0	0
Ambati Rayudu	124.5	0	30	12	3
Abhishek Sharma	46	2	6	3	1
Mahipal Lomror	15.5	0	2	3	0
Alex Carey	8.5	0	0	1	2
Andrew Tye	12	1	0	1	0
Rohit Sharma	149	0	27	19	6
Jimmy Neesham	39	2	0	1	1
Monu Kumar	4	0	0	0	0

d) Statistical Analysis using Multiple Regression

-Here, 4 predictors(Independent variables) and 25 observations are In this model .
So, **k=4 and n=25**

	Coefficientsandard Erro		t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	7.643152	2.927519	2.610795	0.016731	1.536455	13.74985	1.536455	13.74985
X1	9.381209	0.510958	18.36005	5.49E-14	8.31537	10.44705	8.31537	10.44705
X2	1.847696	0.371844	4.96901	7.38E-05	1.072043	2.623349	1.072043	2.623349
X3	3.267733	0.55093	5.931305	8.42E-06	2.118514	4.416953	2.118514	4.416953
X4	4.161098	1.106809	3.759544	0.001234	1.852334	6.469862	1.852334	6.469862

-The above table shows Excel result for the score awarded data multiple regression model. From the table, the computed values of the three regression coefficients are,

b_0: 7.643152
b_1: 9.381209
b_2: 1.847696
b_3: 3.267733
b_4: 4.161098

Therefore, Multiple linear regression equation is,

$$\hat{Y}_i = 7.643152 + 9.381209\, X_{1i} + 1.847696\, X_{2i} + 3.267733 X_{3i} + 4.161098 X_{4i}$$

where,

$\hat{Y}_i$: predicted score of player ί
X_{1i} : no. of wicket taken by player ί
X_{2i} : no. of four by player ί
X_{3i} : no. of six by player ί
X_{4i} : no. of catch by player ί

We can use the multiple regression equation to predict values of the dependent variable.Using the multiple regression equation with $X_{1i} = 7$, $X_{2i} = 4$, $X_{3i} = 6$ **and** $X_{4i} = 5$,

$$\hat{Y}_i = 7.643152 + 9.381209(7) + 1.847696(4) + 3.267733(6) + 4.161098(5)$$
$$\hat{Y}_i = 121.1143$$

Thus,121.1143 **t**he predicted score of one player for 7 wickets taken by him , 4 fours ,6 sixes and 5 catches during IPL 2017 .

Interpret intercept and regression coefficients:

In study of score prediction data,

-The sample Y intercept estimates the number of scores awarded in a match if the wickets taken is 0, Four is 0, Six is 0 and no. of catches is also 0. These values of wickets,fours,sixes and catch are inside the range of wickets,fours,sixes and catch used in the test-market study, So predicted score would be **7.643152** they make sense in the context of the problem, This value because of other factors like wide ball,homeground advantages or previous performance.

-For a player with constant no. of fours, sixes and catches, the estimated score is predicted to increase by **9.381209** per match for each 1 unit increase in the no of wickets taken. Another way to interpret this “net effect” is to think of two players with an equal no. of fours, sixes and catches. If the first player creates 1 unit more than the other player , the net effect of this difference is that the first player is predicted to award **9.381209** more score per match than the second player.
-For a player with constant no. of wickets, sixes and catches, the estimated score is predicted to increase by **1.847696** per match for each 1 unit increase in the no of fours. Another way to interpret this “net effect” is to think of two players with an equal no. of wickets, sixes and catches. If the first player creates 1 unit more than the other player , the net effect of this difference is that the first player is predicted to award **1.847696** more score per match than the second player.
-For a player with constant no. of wickets, fours and catches, the estimated score is predicted to increase by **3.267733** per match for each 1 unit increase in the no of sixes. Another way to interpret this “net effect” is to think of two players with an equal no. of wickets, fours and catches. If the first player creates 1 unit more than the other player , the net effect of this difference is that the first player is predicted to award **3.267733** more score per match than the second player.
-For a player with constant no. of wickets, fours and sixes, the estimated score is predicted to increase by **4.161098** per match for each 1 unit increase in the no of catches. Another way to interpret this “net effect” is to think of two players with an equal no. of wickets, fours and sixes. If the first player creates 1 unit more than the other player , the net effect of this difference is that the first player is predicted to award **4.161098** more score per match than the second player.
-In short, 1 unit increase in wicket is predicted to increase score by **9.381209**, with a fixed no. of fours, sixes and catches.1 unit increase in four is predicted to increase score by **1.847696**, with a fixed no. of wickets, sixes and catches.1 unit increase in six is predicted to increase score by **3.267733**, with a fixed no. of wickets, fours and catches.1 unit increase in catch is predicted to increase score by **4.161098**, with a fixed no. of wickets, fours and sixes.

e) Findings/Conclusions

Interpreting Regression Statistics:

Regression Statistics	
Multiple R	0.989733
R Square	0.979572
Adjusted R	0.975486
Standard E	9.649521
Observatic	25

-The multiple correlation coefficient (= 0.9897) indicates that the scores are strongly correlated with wickets,fours,sixes and catches taken by a player

-The coefficient of multiple determination (= 0.9796) indicates that 97.96% of the variation Score is explained by the variation in wickets,fours,sixes and catches taken by a player.

-The adjusted coefficient of multiple determination (= 0.9755) indicates that 97.55% of the variation in score is explained by the multiple regression model adjusted for the number of independent variables and the sample size.

-The standard error of an estimate (= 9.6495) indicates that the predicted scores are differed by 10 units from the actual scores. Approximately, 84 % (4 out of 25) **residuals** are within the interval (−9.6495 , 9.6495).This indicates that the multiple regression model is appropriate for predicting score.

Observation	Predicted Y	Residuals	Standard Residuals
1	36.38902	-3.389018112	-0.384732379
2	70.26218	31.23782343	3.5462195
3	131.1762	-0.676233212	-0.076768198
4	7.643152	-1.643151893	-0.186535957
5	98.56852	4.931479549	0.559837627
6	95.11789	-8.617889083	-0.978330849
7	113.3272	1.672830049	0.189905118
8	122.854	3.645962854	0.413901583
9	7.643152	0.356848107	0.04051056
10	91.6054	3.394604851	0.385366604
11	160.9612	1.538825131	0.174692443
12	46.911	3.588998343	0.407434786
13	52.99536	-5.995363753	-0.680613228
14	15.03394	-5.033936456	-0.571468869
15	27.54015	-1.540150179	-0.174842867
16	251.6445	-15.1444692	-1.71924949
17	18.87206	6.127943157	0.695664074
18	114.7701	9.729873544	1.104566948
19	51.45604	-5.45604345	-0.61938783
20	21.14174	-5.641743236	-0.640469075
21	19.23308	-10.73308098	-1.218454325
22	20.29209	-8.292093723	-0.941345498
23	144.5845	4.415536724	0.501266116
24	33.8344	5.165599434	0.586415679
25	7.643152	-3.643151893	-0.413582474

Correlation matrix:

	Y	X1	X2	X3	X4
Y	1				
X1	0.579492	1			
X2	0.761668	-0.00955	1		
X3	0.74131	-0.03275	0.855093	1	
X4	0.657585	0.035685	0.70134	0.633086	1

-The correlation coefficient for score and wicket(=0.5795) indicates that scores are positively correlated to wickets with moderate strength above average.

-The correlation coefficient for score and four(=0.7617) indicates that scores are positively correlated to four with strong strength.

-The correlation coefficient for score and six(=0.7413) indicates that scores are positively correlated to six with moderate strength above average.

-The correlation coefficient for score and catch(=0.6579) indicates that scores are positively correlated to catch with moderate strength above average.
-The correlation coefficient for wicket and four(=-0.00955) indicates that wickets are negatively correlated to four with very poor strength.
-The correlation coefficient for wicket and six(=-0.03275) indicates that wickets are negatively correlated to six with very poor strength.
-The correlation coefficient for wicket and catch(=0.03569) indicates that wickets are positively correlated to catch with very poor strength.
-The correlation coefficient for four and six(=0.85509) indicates that fours are positively correlated to six with strong strength.
-The correlation coefficient for four and catch(=0.70134) indicates that fours are positively correlated to catch with moderate strength above average.
-The correlation coefficient for six and catch(=0.6331) indicates that six are positively correlated to six with moderate strength above saverage .

Multicollinearity :

Above Value of Correlation coefficient(R) between Independent variable
=> Among R_{X1X2} , R_{X1X3} , R_{X1X4} , R_{X2X3} , R_{X2X4} , R_{X3X4} . Only biggest value $R_{X2X3} = 0.8551 > 0.75$,
So,Checking multicollinearity first as of X2vsX1X3X4

Regression Statistics	
Multiple R	0.879724
R Square	0.773915
Adjusted R	0.741617
Standard E	5.662855
Observatic	25

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	3	2305.214	768.4045	23.96178	5.59E-07
Residual	21	673.4264	32.06792		
Total	24	2978.64			

X2vsX3X4X1

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	1.931034	1.665545	1.1594	0.259307	-1.53266	5.394725	-1.53266	5.394725
X3	1.102791	0.215919	5.107437	4.65E-05	0.653763	1.551818	0.653763	1.551818
X4	1.182751	0.596055	1.984298	0.060444	-0.05681	2.422314	-0.05681	2.422314
X1	0.009824	0.29985	0.032762	0.974174	-0.61375	0.633395	-0.61375	0.633395

Here $\mathbf{R_{X2}^2 = 0.7739 > 0.75}$ and $\mathbf{VIF = 1/(1-R_{X2}^2) = 4.42}$ which belongs to [4,10]. This is a condition to check Multicollinearity of Independent variables.
-So, We can conclude that there is a **potential problem of Multicollinearity** in the model due to X2.
-Other models has **No problem with multicollinearity** which I've checked due to $R^2 < 0.75$ shown below.

Regression Statistics	
Multiple R	0.080226
R Square	0.006436
Adjusted R	-0.1355
Standard E	4.121082
Observatic	25

ANOVA

	df	SS	MS	F	ignificance F
Regressior	3	2.310371	0.770124	0.045346	0.986811
Residual	21	356.6496	16.98332		
Total	24	358.96			

X1vsX2X3X4

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	1.867004	1.182032	1.579488	0.12917	-0.59117	4.325174	-0.59117	4.325174
X2	0.005203	0.158802	0.032762	0.974174	-0.32504	0.335449	-0.32504	0.335449
X3	-0.05725	0.234957	-0.24368	0.809846	-0.54587	0.431367	-0.54587	0.431367
X4	0.138807	0.471721	0.294256	0.771451	-0.84219	1.119803	-0.84219	1.119803

Regression Statistics	
Multiple R	0.856812
R Square	0.734128
Adjusted R	0.696146
Standard E	3.822081
Observatic	25

ANOVA

	df	SS	MS	F	ignificance F
Regressior	3	847.0657	282.3552	19.32841	3E-06
Residual	21	306.7743	14.6083		
Total	24	1153.84			

X3vsX1X2X4

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-0.50274	1.15436	-0.43551	0.667635	-2.90336	1.897887	-2.90336	1.897887
X1	-0.04925	0.2021	-0.24368	0.809846	-0.46954	0.371043	-0.46954	0.371043
X2	0.502368	0.09836	5.107437	4.65E-05	0.297817	0.706919	0.297817	0.706919
X4	0.187628	0.43648	0.429865	0.671673	-0.72008	1.095338	-0.72008	1.095338

Regression Statistics									
Multiple R	0.705755								
R Square	0.498091								
Adjusted R	0.426389								
Standard E	1.902494								
Observatic	25								
ANOVA									
	df	SS	MS	F	ignificance F				
Regressior	3	75.43085	25.14362	6.946741	0.002001				
Residual	21	76.00915	3.619484						
Total	24	151.44				X4vsX1X2X3			
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%	
Intercept	0.682314	0.557653	1.223547	0.234675	-0.47739	1.842017	-0.47739	1.842017	
X1	0.029582	0.100533	0.294256	0.771451	-0.17949	0.238652	-0.17949	0.238652	
X2	0.133496	0.067276	1.984298	0.060444	-0.00641	0.273405	-0.00641	0.273405	
X3	0.046488	0.108146	0.429865	0.671673	-0.17841	0.271391	-0.17841	0.271391	

Test for the Significance of the Overall Multiple Regression Model:

ANOVA								
	df	SS	MS	F	ignificance F			
Regressior	4	89300.19	22325.05	239.7623	1.37E-16			
Residual	20	1862.265	93.11326			because of, F/P value < 0.05		
Total	24	91162.46						
				SIGNIFICANT				

-The overall significance of the multiple regression model is tested with the following hypotheses:

H_0: $\beta_1 = \beta_2 = 0$

H_a: Atleast one regression coefficient is not equal to 0.

From the ANOVA Table, it can be seen that the significance F value is **(1.37e-16)=0**. At 5% significance level, $\alpha = 0.05$, the significance F value is smaller than 0.05. This indicates that the null hypothesis H_0 will be rejected and we conclude that at least one of the independent variables (no. of wickets,fours,sixess and/or catch) is significantly related to score. Hence, we can say that the overall developed regression model is **statistically significant** and can be used for prediction purpose.

Test for the Significance of the population regression coefficients :

ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	4	89300.19	22325.05	239.7623	1.37E-16
Residual	20	1862.265	93.11326		
Total	24	91162.46			

because of, F/P value < 0.05

SIGNIFICANT

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	7.643152	2.927519	2.610795	0.016731	1.536455	13.74985	1.536455	13.74985
X1	9.381209	0.510958	18.36005	5.49E-14	8.31537	10.44705	8.31537	10.44705
X2	1.847696	0.371844	4.96901	7.38E-05	1.072043	2.623349	1.072043	2.623349
X3	3.267733	0.55093	5.931305	8.42E-06	2.118514	4.416953	2.118514	4.416953
X4	4.161098	1.106809	3.759544	0.001234	1.852334	6.469862	1.852334	6.469862

-The significance of the slope coefficient for wicket can be examined with the following hypotheses:
H_0: $\beta_1 = 0$
and
H_a: $\beta_1 \neq 0$.
From the t-test Table, it can be seen that …
The p-value for wicket is **(5.49e-14)=0**. At 5% significance level, $\alpha = 0.05$, the p-value is smaller than 0.05. This indicates that the null hypothesis H_0 will be rejected and we conclude that there is a significant relationship between the wicket and score, taking into account Four,Six,catch. Hence, we can say that the wicket is **statistically significant** and should be included in the developed regression model.

-The significance of the slope coefficient for four can be examined with the following hypotheses:
H_0: $\beta_2 = 0$
and
H_a: $\beta_2 \neq 0$.
From the t-test Table, it can be seen that …
The p-value for four is **(7.38e-05)=0**. At 5% significance level, $\alpha = 0.05$, the p-value is smaller than 0.05. This indicates that the null hypothesis H_0 will be rejected and we conclude that there is a significant relationship between the four and score, taking into account wicket,Six,catch. Hence, we can say that the four are **statistically significant** and should be included in the developed regression model.

-The significance of the slope coefficient for six can be examined with the following hypotheses:
H_0: $\beta_3 = 0$
and

H_a: $\beta_3 \neq 0$.
From the t-test Table, it can be seen that …
The p-value for six is **(8.42e-06)=0**. At 5% significance level, $\alpha = 0.05$, the p-value is smaller than 0.05. This indicates that the null hypothesis H_0 will be rejected and we conclude that there is a significant relationship between the six and score, taking into account Four,wicket,catch. Hence, we can say that the six is **statistically significant** and should be included in the developed regression model.

-The significance of the slope coefficient for catch can be examined with the following hypotheses:
H_0: $\beta_4 = 0$
and
H_a: $\beta_4 \neq 0$.
From the t-test Table, it can be seen that …
The p-value for catch is **0.001234**. At 5% significance level, $\alpha = 0.05$, the p-value is smaller than 0.05. This indicates that the null hypothesis H_0 will be rejected and we conclude that there is a significant relationship between the catch and score, taking into account Four,Six,wicket. Hence, we can say that the catch is **statistically significant** and should be included in the developed regression model.

Confidence Interval Estimation:

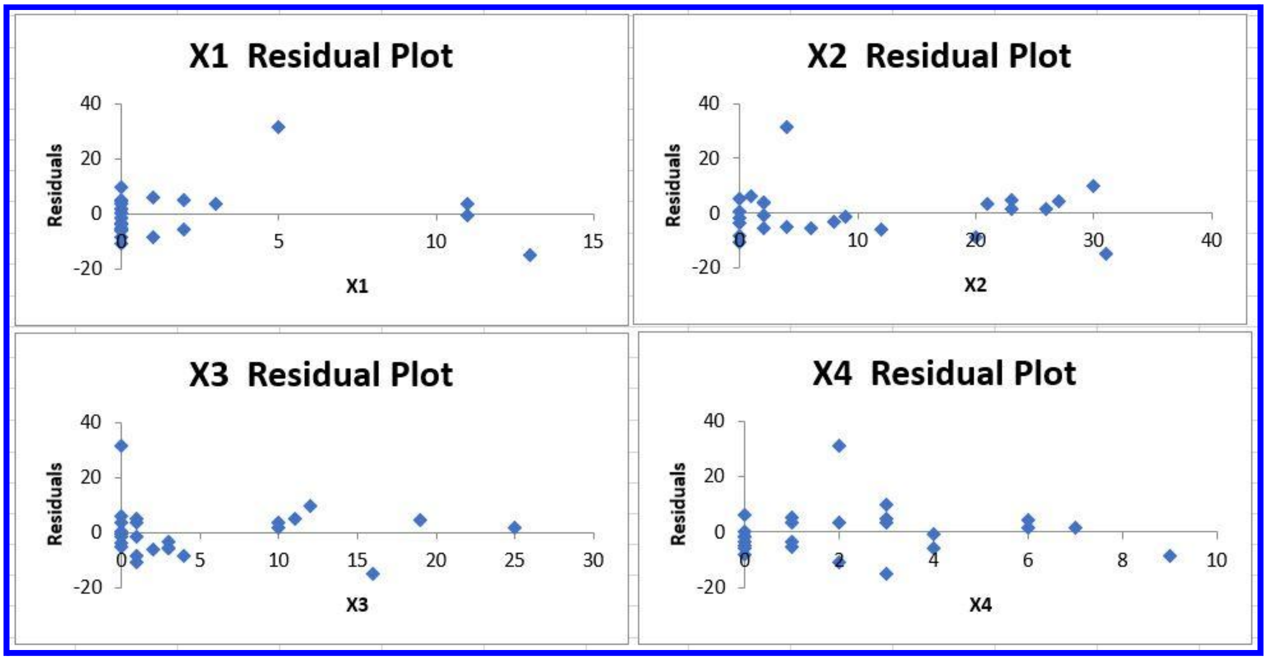
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	7.643152	2.927519	2.610795	0.016731	1.536455	13.74985	1.536455	13.74985
X1	9.381209	0.510958	18.36005	5.49E-14	8.31537	10.44705	8.31537	10.44705
X2	1.847696	0.371844	4.96901	7.38E-05	1.072043	2.623349	1.072043	2.623349
X3	3.267733	0.55093	5.931305	8.42E-06	2.118514	4.416953	2.118514	4.416953
X4	4.161098	1.106809	3.759544	0.001234	1.852334	6.469862	1.852334	6.469862

-Taking into account the effect of four,six and catch, the estimated effect of a 1 unit increase in wicket is to increase the mean score by approximately 8.3 to 10.4 score. We have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient has a **significant effect.**
-Taking into account the effect of wicket,six and catch, the estimated effect of a 1 unit increase in four is to increase the mean score by approximately 1 to 2.6 score. We have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient has a **significant effect.**
-Taking into account the effect of wicket,four and catch, the estimated effect of a 1 unit increase in six is to increase the mean score by approximately 2.1 to 4.4 score. We have 95% confidence that this interval correctly estimates the relationship between these variables. From a

hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient has a **significant effect.**

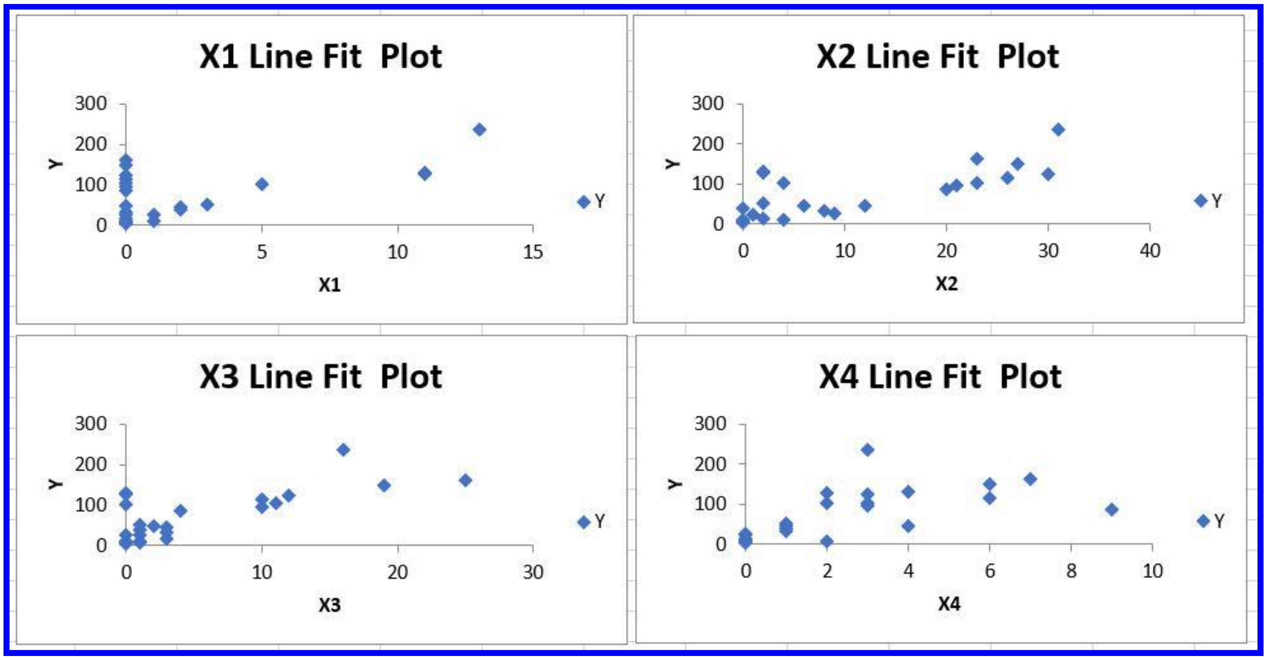
-Taking into account the effect of wicket,four and six , the estimated effect of a 1 unit increase in catch is to increase the mean score by approximately 1.85 to 6.47 score. We have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient has a **significant effect.**

Analyze the Residual plots:



There is very little or **no pattern** in the relationship between the residuals and X1 (wicket), X2 (four), X3(six) and X4(catch) Thus, we can conclude that the multiple regression model is appropriate for predicting score and plot is a **healthy Residual plots.**

Analyze the Line Fit plots:



-There is very little or no pattern in the relationship between the Y (scores) and X1 (wicket), X2 (four), X3(six) and X4(catch) .

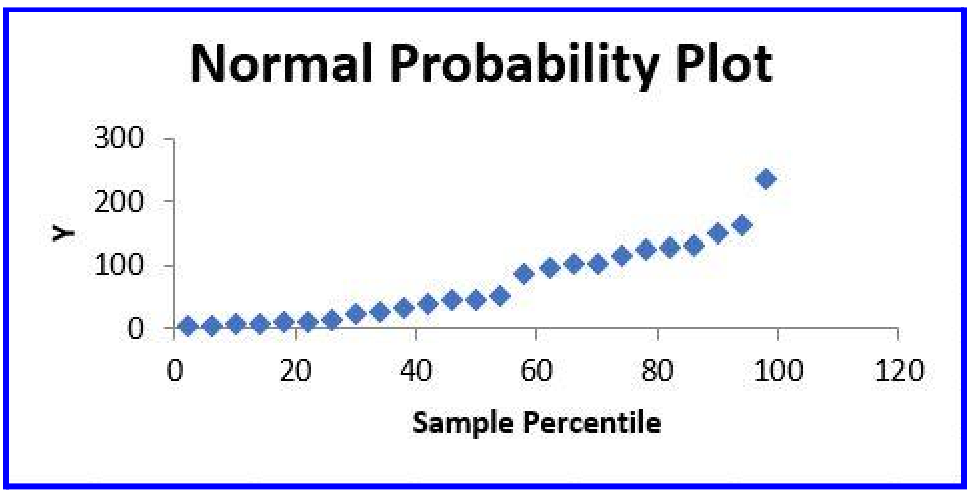
-It can be seen that the Y are **positively** correlated to wickets with moderate strength above average (X1) , the Y are **positively** correlated to four with strong strength(X2) , the Y are **positively** correlated to six with moderate strength above average(X3) and the Y are **positively** correlated to catch with moderate strength above average(X4) .

Detection of outliers:

-The **standard residuals** of the **2nd** player is not within the interval [− 2, 2]. So, according to the thumb rule, these stores may be considered as **outliers**.

Observation	Predicted Y	Residuals	Standard Residuals
1	36.38902	-3.389018112	-0.384732379
2	70.26218	31.23782343	3.5462195
3	131.1762	-0.676233212	-0.076768198
4	7.643152	-1.643151893	-0.186535957
5	98.56852	4.931479549	0.559837627
6	95.11789	-8.617889083	-0.978330849
7	113.3272	1.672830049	0.189905118
8	122.854	3.645962854	0.413901583
9	7.643152	0.356848107	0.04051056
10	91.6054	3.394604851	0.385366604
11	160.9612	1.538825131	0.174692443
12	46.911	3.588998343	0.407434786
13	52.99536	-5.995363753	-0.680613228
14	15.03394	-5.033936456	-0.571468869
15	27.54015	-1.540150179	-0.174842867
16	251.6445	-15.1444692	-1.71924949
17	18.87206	6.127943157	0.695664074
18	114.7701	9.729873544	1.104566948
19	51.45604	-5.45604345	-0.61938783
20	21.14174	-5.641743236	-0.640469075
21	19.23308	-10.73308098	-1.218454325
22	20.29209	-8.292093723	-0.941345498
23	144.5845	4.415536724	0.501266116
24	33.8344	5.165599434	0.586415679
25	7.643152	-3.643151893	-0.413582474

Normal Probability Plot:



-Normal Probability plots are used to verify the assumption of normal distribution. The assumption of normality of disturbances is very much needed for the validity of the results for testing of hypothesis, confidence intervals and prediction intervals.

-Some experience and expertise is required to interpret the normal probability plots because the samples taken from a normal distribution will not plot exactly as a straight line:
-Small sample sizes (n ≤45) often produce normal probability plots that deviate substantially from linearity.
-Larger sample sizes (n ≥95) produces plots which are much better behaved.
-Usually about n = 80 is required to produce stable and easily interpretable normal probability plots.

Best model for prediction:backward elimination method:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	7.643152	2.927518761	2.610795187	0.016731	1.536455	13.74985	1.536455	13.74985
X1	9.381209	0.510957627	18.36005239	5.49E-14	8.31537	10.44705	8.31537	10.44705
X2	1.847696	0.371843948	4.969009584	7.38E-05	1.072043	2.623349	1.072043	2.623349
X3	3.267733	0.550929828	5.931305326	8.42E-06	2.118514	4.416953	2.118514	4.416953
X4	4.161098	1.106809368	3.759543561	0.001234	1.852334	6.469862	1.852334	6.469862

-In backward elimination method,we have to eliminate the highest p-values variable from the model if and only if that variable is insignificant . But here Four variables are statistically significant so we have to stop searching for insignificant var. And **Best model would be Y vs X1,X2,X3,X4**

f) Bibliography (References, Websites, articles, other materials etc.).

[Predictive Analysis of an IPL Match]

[Predicting outcome of IPL match based on variables]

[IPL Match Prediction based on Powerplay]

[Predicting Outcome of IPL Matches]