The data are in the Excel file named "DS604 Fall24 Data for Excel assignment#1.xlxs"
PROBLEM 1:

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year can be predicted using the ACT test score. The data is provided in the "**Problem 1 data**" worksheet.

Determine which variable should be the independent variable (X) and the dependent variable (Y). Use Excel's regression tool to obtain the regression equation for the problem. You must also show the summary output report.

(Y) Dependent– GPA
(X) Independent – ACT

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.269481803							
R Square	0.072620442							
Adjusted R Sq	0.064761293							
Standard Erro	0.623125037							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	3.5878459	3.5878459	9.240242702	0.0029166			
Residual	118	45.8176078	0.38828481					
Total	119	49.4054537						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.11	0.32089483	6.58798167	1.30445E-09	1.47859015	2.74950842	1.47859015	2.74950842
ACT test score	0.04	0.01277302	3.03977675	0.002916604	0.01353307	0.06412118	0.01353307	0.06412118

1.1. What is the regression equation for the problem?

Equation: Y = B0 + B1*(x) = 2.11 + 0.04*(x)

1.2. Comment on the strength of the relationships between X and Y. Is X a good predictor of Y? Explain.

The correlation coefficient helps us understand the relationship between "x" and "y". Looking at the output report, this value is 0.27, indicating that a weak linear relationship exists between the two variables. Using our r-squared value generated by the regression report, we can determine how well "x" predicts "y". The value generated by the report is 0.073, indicating that the ACT score is a poor predictor of GPA.

For Questions 1.3 - 1.5, don't just provide the final answers. You must show intermediate calculations leading up to the final answers.

1.3. Obtain a point estimate of the mean freshman GPA for students with ACT test scores of **30**.

Using the equation from the output report that is documented in (1.1). We can plug in the ACT test score "30" into the equation and generate the following point estimate.

$$Y = 2.114 + (0.039 * 30)$$

$$Y = 3.28$$

1.4. Construct a 98% confidence interval for the mean freshman GPA for students whose ACT test score is **30**.

t-value = 2.36
y(hat) = 3.28
s[e] = 0.62
(A) 1/n = 0.0083
(B) (X(0) – x(bar))^2 = 27.83
(C) SSxx = 2379.93
(D) Sqrt((A) + ((B)/(C)))

Confidence Interval : y(hat) – (t*s[e]*(D)) <= 3.28 <= y(hat) + (t*s[e]*((D))

Confidence Interval: 3.07 <= 3.28 <= 3.49

1.5. Mary Jones obtained a score of **28** on the entrance test. Construct a 90% prediction interval of her freshman GPA.

t-value = 1.66
y(hat) = 3.20
s[e] = 0.62
(A) 1/n = 0.0083
(B) (X(0) – x(bar))^2 = 10.73
(C) SSxx = 2379.93
(D) Sqrt(1 + (A) + ((B)/(C)))

Confidence Interval : y(hat) – (t*s[e]*(D)) <= 3.28 <= y(hat) + (t*s[e]*((D))

Confidence Interval: 2.16 <= 3.20 <= 4.0

PROBLEM 2: BASEBALL FORECAST

Paul Raymond, a math savvy baseball manager, would like to apply his knowledge in statistics to develop a multiple regression model to forecast pitching performances for starting pitchers for this baseball season. He intended to use the baseball statistics from the last year season to build the model. He understood that the initial variable selection was the most important aspect of developing a regression model. He knew that, if he didn't have good predictor variables, he wouldn't end up with useful predicting equations.
Paul had spent considerable amount of time to download the baseball statistics for starting pitchers from the last year season. He also decided to include only starting pitchers who had pitched at least 100 innings during the last season. The data for the 138 pitchers selected is provided in the **"Problem 2 data.xlsx"** worksheet**.**
Paul decided that Earned Run Average (*ERA*) is the best indicator of performance and so wanted to develop a regression model to predict this variable. He chose the six potential predictor variables as follow,

WHIP: Number of walks plus hits given up per inning pitched
CMD: Command of pitches, the ratio strikeouts/walks
K/9: How many batters a pitcher strikes out per game (nine innings pitched)
HR/9: Opposition home runs per gram (nine innings pitched)
OBA: Opposition batting average
THROWS: Right-handed pitcher (1) or left-handed pitcher (0)

QUESTIONS & TASKS:

Part1: Preliminary analysis:

2.1. Present the correlation matrix. Examine the correlation matrix and comment on the correlations among the variables.

	ERA	*WHIP*	*CMD*	*K/9*	*HR/9*	*OBA*	*THROWS*
ERA	1.00						
WHIP	0.82	1.00					
CMD	-0.46	-0.66	1.00				
K/9	-0.45	-0.50	0.57	1.00			
HR/9	0.56	0.34	-0.18	-0.24	1.00		
OBA	0.82	1.00	-0.67	-0.51	0.34	1.00	
THROWS	0.04	-0.04	0.12	-0.06	0.02	-0.04	1.00

Every value encompasses the R-value (correlation) between various variables both independent and dependent. Anything above 0.8 has a strong correlation between the two variables. 0.5 to 0.7 has a moderate correlation, showing an existing but not as strong relationship between the variables. Anything below 0.5 has a weak and/or non-apparent correlation.

Looking at the correlation amongst the dependent variable (ERA) and the six independent variables (WHIP, CMD, K/9, HR/9, OBA, THROWS):

A) Dependent Correlation Analysis

- *WHIP* and *OBA* are strongly correlated to *ERA*.
- *HR/9* is moderately correlated to ERA.
- *CMD*, *K/9*, and *THROWS* are weakly correlated to *ERA*

B) Independent Correlation Analysis

- (1) WHIP:
    - Perfectly correlated with OBA.
    - Moderately correlated CMD and K/9.
    - Weakly correlated w/ THROWS and HR/9.
- (2) OBA:
    - Perfectly correlated with *WHIP*
    - moderately correlated w/ *CMD* and *K/9*
    - weakly correlated w/ *HR/9* and *THROWS*
- (3) CMD:
    - Moderately correlated w/ *WHIP, OBA, and K/9.*
    - Weakly correlated w/ HR/9 and THROWS.
- (4) K/9:
    - Moderately correlated w*/ WHIP, CMD,* and *OBA*
    - Weakly correlated w/ HR/9 and Throws.
- (5) HR/9:
    - *HR/9* is weakly correlated with all variables
- (6) THROWS:
    - THROWS is weakly correlated w/ all variables

2.2. Determine which predictor variables should be used to predict *ERA*. Explain why?

Predictor variables that can be

I believe we only need to use two predictor variables to help predict ERA. These variables are: (1) OBA or WHIP and (2) HR/9. OBA and WHIP can be used interchangeably since they are so similar in strength and correlation. The only weakness here is that both show moderate correlation amongst the independent variables. HR/9 is useful due to its moderate correlation to ERA and weak correlation found amongst independent variables

Part 2: Model development and evaluations: develop the full model and perform statistical tests

2.3. Develop the multiple regression analysis using all variables (i.e. the full model). Present the regression equation and the Excel output report.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.885620399
R Square	0.784323491
Adjusted R Sq	0.774445177
Standard Erro	0.438941763
Observations	138

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	91.7862186	15.2977031	79.3985226	3.2193E-41
Residual	131	25.2397531	0.19266987		
Total	137	117.025972			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.179	2.87242541	-1.802927	0.07369837	-10.861116	0.50356921	-10.861116	0.50356921
WHIP	0.546	4.1534737	0.13156981	0.89552642	-7.67009	8.7630335	-7.67009	8.7630335
CMD	0.124	0.0476707	2.60196639	0.01033603	0.02973354	0.21834156	0.02973354	0.21834156
K/9	-0.030	0.0299201	-1.0185596	0.31028951	-0.0896645	0.02871369	-0.0896645	0.02871369
HR/9	0.841	0.12024459	6.99034076	1.2571E-10	0.60267817	1.07842313	0.60267817	1.07842313
OBA	0.025	0.02658327	0.92177546	0.35834021	-0.0280842	0.07709184	-0.0280842	0.07709184
THROWS	0.078	0.08538991	0.91378185	0.36251026	-0.0908939	0.24694935	-0.0908939	0.24694935

Equation: $Y = B(0) + B(1)*X(1) + B(2)*X(2) + B(3)*X(3) + B(4)*X(4) + B(5)*X(5) + B(6)*X(6)$

$-5.179 + 0.546*(WHIP) + 0.124*(CMD) + -0.030*(K/9) + 0.841*(HR/9) + 0.025*(OBA) + 0.078*(THROWS)$

2.4. Conduct the t-test (using α = 0.01) to determine which predictor variables, if any, are significant. Provide the details of the test and comment on the results of the test.

Variable	t		CV	
WHIP	1.80292704	<	2.61	accept h(0)
CMD	0.13156981	<	2.61	accept h(0)
K/9	2.60196639	<	2.61	accept h(0)
HR/9	1.01855961	<	2.61	accept h(0)
OBA	6.99034076	>	2.61	reject h(0)
THROWS	0.92177546	<	2.61	accept h(0)

Looking at the table, all t-values values that are not greater than the Critical Value of 2.61 are labeled in red, those that are greater than the Critical T-Value are labeled in Green. The conclusions we can draw from this table are:

- The observed t-score for the OBA variable is greater than our Critical Value, meaning that we can reject our null hypothesis H(0) of no correlation and accept are alternative hypothesis H(A) of existing correlation. This treats the OBA variable as significant
- All other variables have an observed t-score less than our Critical T-Value, meaning that we must accept our null hypothesis of no correlation for these individual variables and reject our alternative hypothesis of existing correlation. This treats all variables other than OBA as insignificant

2.5. Conduct the F-test (using α = 0.01) to test the overall significance of the model. Provide the details of the test and comment on the results of the test.

F		CV	
79.4	>	2.94	reject H(0)

The F-Value generated by the output report is greater than our critical F-value. This means that we can reject the null hypothesis that there is no correlation amongst at least one variable and accept

our alternative hypothesis that there exists correlation amongst at least one variable. This conclusion helps accept our overall model as significant

2.6. Comment on the values of r^2 and adjust r^2

The r-squared value from the output report is 0.78 and the adjusted r-squared comes out to 0.77. This indicates “x” is a relatively good predictor of “y”.

2.7. Based on the results of your tests, if you want to drop more predictor variable(s) to simplify the model, which one(s) should be dropped? Explain in detail.

After the following tests, I believe that the only predictor variable that should be included is OBA. Since OBA is the only variable to pass the t-test, this will help in simplifying the model down to a single variable and maintaining significant correlation. Using the F-test further confirms this in its definitive nature by proving that at least one of the variables has correlation (OBA).

Part 3: Simplified model:

2.8. Drop the predictor variable(s) you recommended in Part 2. Develop a simplified regression model using the remaining predictor variables. Present the regression equation of the simplified model and Excel output report.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8246881
R Square	0.68011046
Adjusted R Sq	0.67775833
Standard Erro	0.52465244
Observations	138

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	79.5905875	79.5905875	289.146755	1.8056E-35
Residual	136	37.4353843	0.27526018		
Total	137	117.025972			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-4.67416	0.54242238	-8.6171961	1.5436E-14	-5.7468333	-3.6014868	-5.7468333	-3.6014868
OBA	0.02873084	0.00168962	17.0043158	1.8056E-35	0.02538951	0.03207217	0.02538951	0.03207217

Equation: Y = B(0) + B(1) * X = -4.67 + 0.029 * (OBA)

2.9. Evaluate the simplified model using the same statistical analyses in Part 2.

r-squared: 0.68

adjusted r-squared: 0.68

t-test: (significant)

Variable	t		CV	
OBA	17.0043158	>	2.61	reject h(0)

F-test: (significant)

F		CV	
289.146755	>	6.82	reject H(0)

2.10. Write the summary of your findings and conclusions.

After simplifying the model down to one predictor variable we can maintain the correlative significance between only one independent variable OBA and the dependent variable ERA. Even though both our r-squared values only show the strength of moderate correlation, both significance tests pass, individual (t) and overall (F), reaffirming the significance of OBA's correlation with ERA when compared to other variables that otherwise do not pass under similar alphas equal to 0.01. This helps simplify data collection and enables the Baseball manager Paul Raymond to only focus on predictive variables that are significantly correlated to ERA. By focusing on this variable alone and noting this relationship is positively correlated, Paul can try to minimize the oppositions batting average to thus decrease their Earned runs average.