Data Science Individual Assessment

The purpose of this assessment is to understand how balance on credit cards relates to various characteristics of the user (income, age, years of education, gender, whether a student or not, marital status, ethnicity, and credit rating) as well as of the credit cards themselves (total credit limit and the number of cards, in addition to balance). Assume that the currency is US dollars (assume income is in thousand of US dollars per year).

1. **A company manager says that the average balance on their credit cards is $500. Do you think that this assertion is justified? Use a one-sample t-test to draw your conclusion**.

   Ans:- We got the following observations from this average mean & the shadow variable.

	Balance	*Dummy*
Mean	520.015	0
Variance	211378.2253	0
Observations	400	3
Hypothesized Mean Difference	500	
df	399	
t Stat	0.870673781	
P(T<=t) one-tail	0.192227914	
t Critical one-tail	1.648681534	
P(T<=t) two-tail	0.384455827	
t Critical two-tail	1.965927296	

To calculate one sample t-test from our given dataset we found credit balance is $500 & Dummy variable is 0 which is ignored also.
we got one-tail P-value is 0.192227914 & Hypothesized Mean Difference is 500 .

As of we know that
The null hypothesis says that a population parameter(mean, standard deviation) is equal to a hypothesized value. and if one-tail p-value is greater than 0.05(typically greater than or equal to 0.05) then we haven't enough evidence to reject null hypothesis & accept the alternative hypothesis.

The alternative hypothesis states that **a** population parameter is smaller, greater, or different than the hypothesized value in the null hypothesis And if one-tail p-value is less than 0.05 then we have evidence to reject the null hypothesis & accept the alternative hypothesis.

in this case from the **one-tail sample t-test** our p-value (0.192227914) is greater than 0.05 so we have not enough evidence to reject the null hypothesis & accept the alternative hypothesis.

So from the one tail sample t-test we got the conclusion that the **null hypothesis can't be rejected(fail to reject the null hyphothesis)** & the manager statement is **true** that the average balance on their credit card is $500.

*Calculation part in tab(answer-1) in excel sheet.

2. **Is there a difference between men and women as far as average balance is concerned? Use a two-sample t-test to draw your conclusion.**

From our dataset we have calculated two sample t-test & got the following output.

t-Test: Two-Sample Assuming Unequal Variances

	Male Balance	*Female Balance*
Mean	509.8031088	529.5362319
Variance	213554.5652	210187.1043
Observations	193	207
Hypothesized Mean Difference	0	
Df	396	
t Stat	-0.42838443	
P(T<=t) one-tail	0.334302083	
t Critical one-tail	1.648710601	
P(T<=t) two-tail	0.668604165	
t Critical two-tail	1.965972608	

If the two means are to be compared, the hypothesis test compared the population means. The null hypothesis is H0: mean1=mean2 & the alternative hypothesis can be two sided or one sided. H1: mean1 not equal to mean2, H1:mean1 greater than mean2 &H1:mean1 less than mean2.

If one-tail p-value is greater than 0.05(typically greater than or equal to 0.05) then we haven't enough evidence to reject null hypothesis & accept the alternative hypothesis.

And if one-tail p-value is less than 0.05 then we have evidence to reject the null hypothesis & accept the alternative hypothesis.

in our case from the **two-tailed sample t-test** our p-value (0.668604165) is greater than 0.05 so we have not enough evidence to reject the null hypothesis & accept the alternative hypothesis. Fail to reject the null hypothesis.

In our case from the two-tailed sample t-test :-
H0:Population Average Balance is the same for Male & Female.
H1:Population Average Balance not is the same for Male & Female.
H1:Population Average Balance is less in Male than female.

p-value(0.668604165424214) greater than 0.05.
we fail to reject the Ho,null hypothesis in favour of alternative hyphothesis or H1.

So from the two tailed sample t-test we got the conclusion that the **null hypothesis can't be rejected**. we have not enough evidence to say that average balance depends on men and women.

*Calculation part in tab(answer-2) in excel sheet.

3. **Is there a difference between students and non-students as far as average balance is concerned? Use a two-sample t-test to draw your conclusion**.

	Yes_Student_Balance	*No_Student_Balance*
Mean	876.825	480.3694444
Variance	240101.9429	193085.1361
Observations	40	360
Hypothesized Mean Difference	0	
df	46	
t Stat	4.902778661	
P(T<=t) one-tail	6.08619E-06	
t Critical one-tail	1.678660414	
P(T<=t) two-tail	1.21724E-05	
t Critical two-tail	2.012895599	

If the two means are to be compared, the hypothesis test compared the population means. The null hypothesis is H0: mean1=mean2 & the alternative hypothesis can be two sided or one sided. H1: mean1 not equal to mean2, H1:mean1 greater than mean2 &H1:mean1 less than mean2.

If one-tail p-value is greater than 0.05(typically greater than or equal to 0.05) then we haven't enough evidence to reject null hypothesis & accept the alternative hypothesis.

And if one-tail p-value is less than 0.05 then we have evidence to reject the null hypothesis & accept the alternative hypothesis.

in our case from the **two-tailed sample t-test** our p-value (0.0000121723866277271) is less than 0.05 then we have evidence to reject the null hypothesis & accept the alternative hypothesis.

In our case from the two-tailed sample t-test :-

H0:Population Average Balance is the same for student & non- Students.
H1:Population Average Balance not is the same for student & non -Students. OR
H1:Population Average Balance is greater than student than non-students.

p-value(0.0000121723866277271) less than 0.05.
Reject H0 in favour of H1.

So from the two test sample t-test we got the conclusion that the **null hypothesis is rejected** in favour of alternative hypothesis.
So Enough evidence to say that the population average balance depends on student.

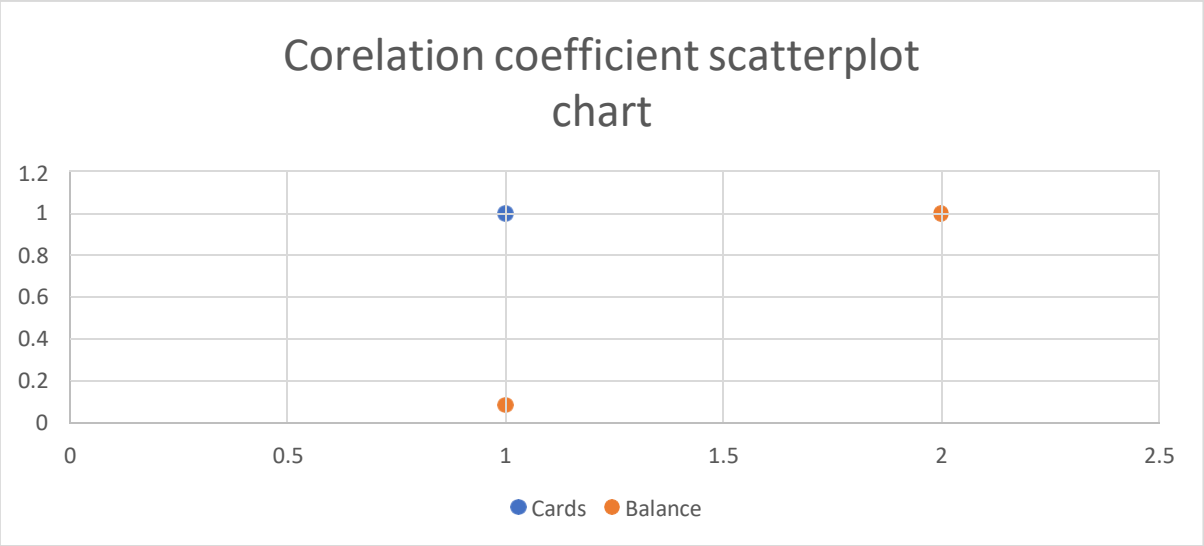
*Calculation part in tab(answer-3) in excel sheet.

4. **It is generally assumed that if there are more credit cards then the balance on the cards will be more. Based on this dataset, do you think this is true? Calculate a correlation coefficient and show a scatter plot to support your answer**.

Data Science Individual Assessment

The co-relation between cards & balance outcome is

	Cards	*Balance*
Cards	1	
Balance	0.086456	1



Statistics Point of View :-

capture the relationship between two continuous quantitative variables and then we can figure out whether those variables have tight linear relationship. because the dots will simply line up along the Likert scale values, rather than being scattered. We found the low co-efficient value value is 0.001835119. There is **no relationship** between Cards & Balance also.

this is false based on our dataset, when more credit cards the balance on the cards will be more.

*Calculation part in tab(answer-4) in excel sheet.

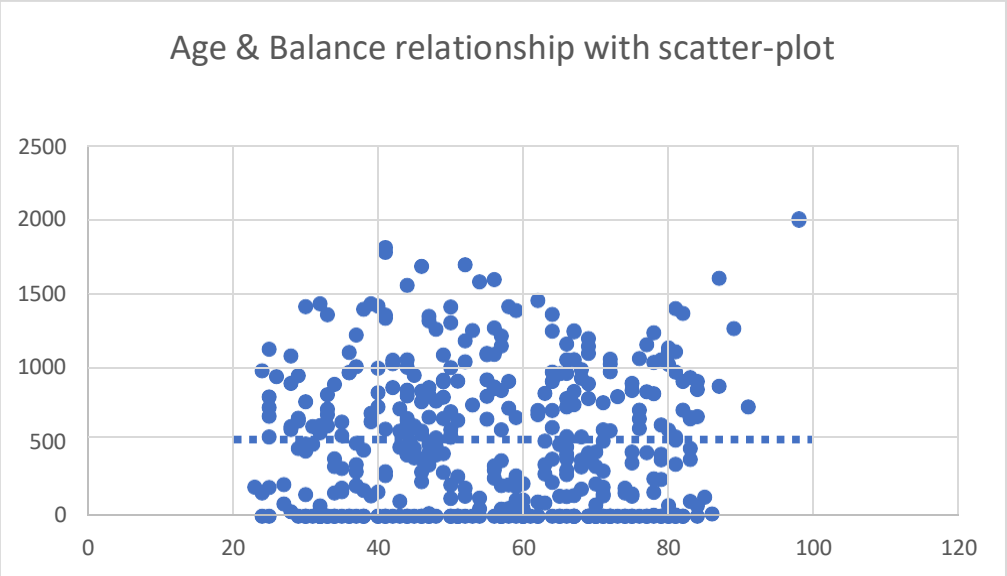
5 Examine whether the following demographic variables influence balance: (a) age, (b) years of education, (c) marital status. For age and years of education, use scatter plots to depict their relationship with balance and calculate the correlation coefficient. For the relationship between marital status and balance, use a two-sample t-test to draw your conclusion

Compare between Age & Balance we got following co-relation value from excel.

	Age	*Balance*
Age	1	
Balance	0.001835119	1

And providing the scatter plot chart with relationship between Age & Balance is given below.

Data Science Individual Assessment



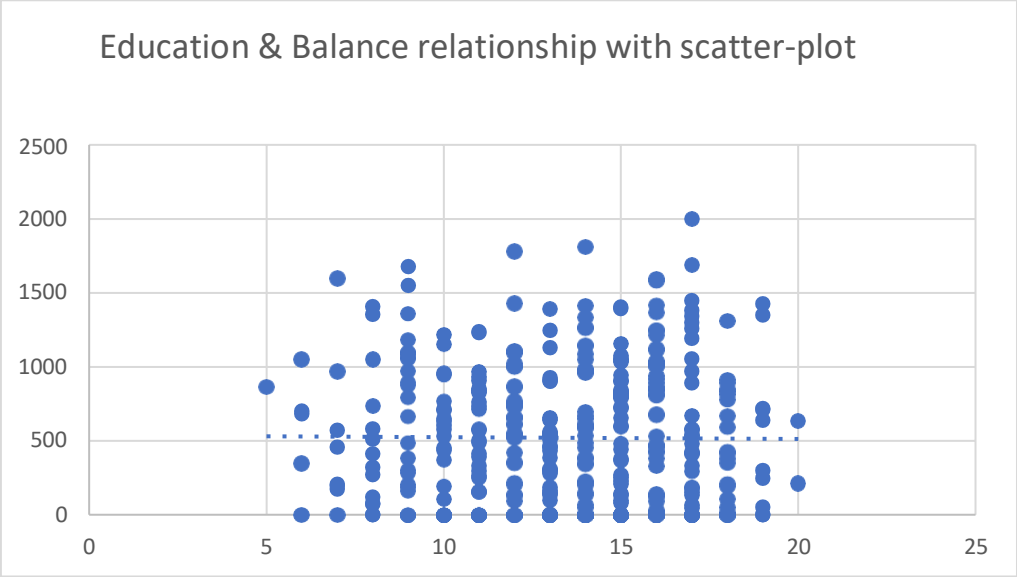
Statistics Point of View :-

co-relation is a number between -1 and +1. Where +1 indicates a perfect positive linear relationship between X and Y. And -1 indicates a perfect negative linear relationship between X and Also 0 indicates either no relationship or a complex nonlinear relationship. We found the low co-effieient value is 0.001835119. There is **no relationship** between Age & Balance. If we observe the trend line it just a straight line, not prove any relationship between age & balance.

	Education	*Balance*
Education	1	
Balance	-0.008061576	1

And providing the scatter plot chart with relationship between Education & Balance is given below.

Data Science Individual Assessment



Statistics Point of View :-

co-relation is a number between -1 and +1. Where +1 indicates a perfect positive linear relationship between X and Y. And -1 indicates a perfect negative linear relationship between X and Also 0 indicates either no relationship or a complex nonlinear relationship. We found the negative co-efficient value is -0.008061576. There is **no relationship** between education & Balance. If we observe the trend line it just a straight line, not prove any relationship between education & balance.

Now comparing the Relationship between Martial Status & Balance, the following two sample t-test drawn.

*Calculation part in tab (answer-5.age&education&balance) in excel sheet.

t-Test: Two-Sample Assuming Unequal Variances

	Married_Yes_Balance	*Married_No_Balance*
Mean	517.9428571	523.2903226
Variance	205696.7262	221735.0385
Observations	245	155
Hypothesized Mean Difference	0	
df	319	
t Stat	-0.112233601	
P(T<=t) one-tail	0.455354389	
t Critical one-tail	1.649644319	
P(T<=t) two-tail	0.910708777	
t Critical two-tail	1.967428387	

Statistics Point of View :-

If the two means are to be compared, the hypothesis test compared the population means. The null hypothesis is H0: mean1=mean2 & the alternative hypothesis can be two sided or one sided. H1: mean1 not equal to mean2, H1:mean1 greater than mean2 &H1:mean1 less than mean2.

If one-tail p-value is greater than 0.05(typically greater than or equal to 0.05) then we haven't enough evidence to reject null hypothesis & accept the alternative hypothesis.

And if one-tail p-value is less than 0.05 then we have evidence to reject the null hypothesis & accept the alternative hypothesis.

in our case from the **two-tailed sample t-test** our p-value (0.910708777) is greater than 0.05 so we have not enough evidence to reject the null hypothesis & accept the alternative hypothesis. Fail to reject the null hypothesis.
And here t-stat is -0.112233601 that simply means that it lies to the left of the mean. All values to the left of the mean are negative and positive to the right of the mean.

In our case from the two-tailed sample t-test :-
H0:Population Average Balance is the same for Married & non-Married.
H1:Population Average Balance not is the same for Married & non-Married.
H1:Population Average Balance is greater in Married than non-Married.
H1:Population Average Balance is less in Male than female.

p-value(0.910708777109252) greater than 0.05..
we fail to reject the null hypothesis in favour of alternative hyphothesis.

So from the two tailed sample t-test we got the conclusion that the **null hypothesis can't be rejected**. There is no relationship between Martial Status & average Balance. we have not enough evidence to say that average balance depends on martial status.

*Calculation part in tab (answer-5. martialstatus&balance) in excel sheet.

6 "Ethnicity of the cardholder does not matter as far a balance is concerned." Carry out an analysis of variance (ANOVA) and discuss whether this statement is supported by the data or not.

Anova: Single Factor(Multiple mean value)

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Ethnicity_Asia_Balance	102	52256	512.3137	231748.3362
Ethnicity_Caucasian_Balance	199	103181	518.4975	190922.4129
Ethnicity_African American_Balance	99	52569	531	235839.1633

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	18454.2	2	9227.1	0.043442783	0.957492	3.018452
Within Groups	84321458	397	212396.6			
Total	84339912	399				

Statistics Point of View :-
From this analysis got following observations like:-

a) Between groups there is variability. Different Averages of each group.
b) There is variability within groups, Different variances of each group.

c) the variance is a measure of how far the observations are from the average.

d) ss sum of squares of the difference from the means. Therefore, this is the measure of total variability. That is decomposed with two groups between & within. Between groups there is variability. This test is called analysis of variance.
e) Essentially variance of mean is computed & compare to 0.
f) H0 is population average balance which is same for all Ethnicity.
g) H1 is population average balance which is not same for all Ethnicity.

h) in our case two-tail p-value(0.957491888119669) is greater than 0.05(typically greater than or equal to 0.05) then we haven't enough evidence to reject null hypothesis & accept the alternative hypothesis. so we fail to reject the null hypothesis in favour of alternative hypothesis. We have not enough evidence to say that average balance depends on Ethnicity of the card holder.

*Calculation part in tab (answer-6(anova) in excel sheet.

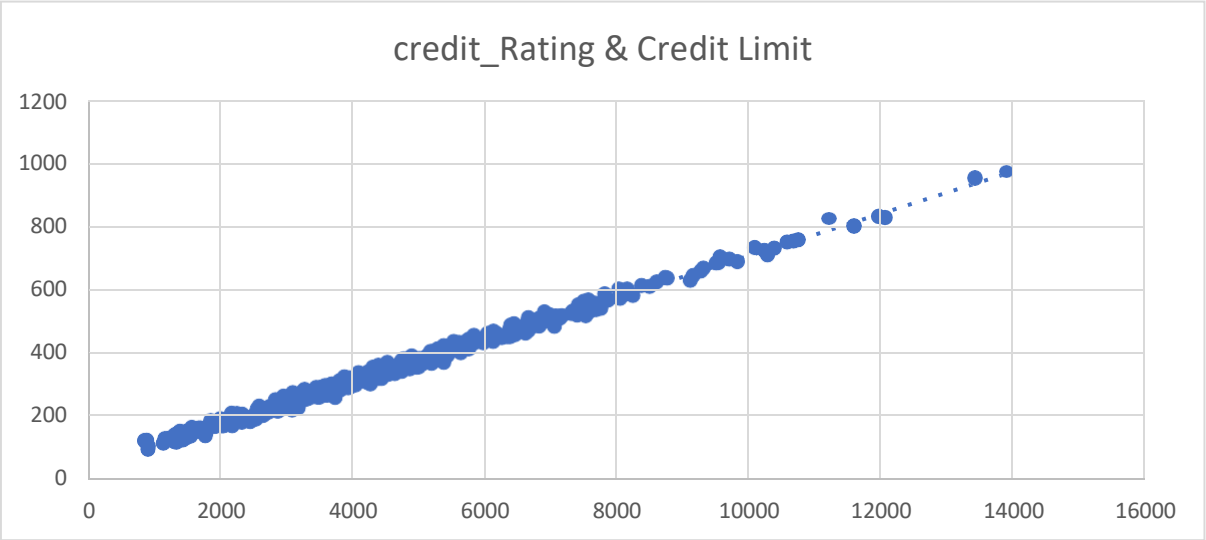
7 A general principle that credit card companies often follow is to assign a higher credit limit to people with a higher credit rating. Does the data show that this principle is being followed?

Statistics Point of View :-
From this analysis got observations for co-relation between credit rating & credit limit. The co-relation is showing in tabular format here along with scatter plot chart.

	credit Limit	*credit Rating*

Data Science Individual Assessment

credit Limit	1	
credit Rating	0.996879737	1



So we can provide the conculation in following ways.

1) co-relation is a number between -1 and +1.
2) "+1" indicates a perfect positive linear relationship between X and Y.
3) "-1" indicates a perfect negative linear relationship between X and Y.
4) 0 indicates either no relationship or a complex nonlinear relationship.
5) The co-efficient value for towards one or mostly 1
6) there is strong relationship between credit limit & rating. And the credit rating value is towards the one, positive relationship.

   Yes the credit card companies often follow is to assign a higher credit limit to people with a higher credit rating.

*Calculation part in tab (answer-7) in excel sheet.

8 Run a simple linear regression of balance on the credit limit. (Here credit limit is the X and the balance is the Y). Report the coefficients and the R-squared. Show a scatter plot.

Ans:- we got the following table of results from this analysis.
Statistics Point of View:- (y= a+bx)

1) P- value is less than 0.05 so coefficient is significantly different from zero.
2) R-squared signifies the coefficient of determination which shows the goodness of fit.it shows how many points fall on the regression line. In our case the value of R-square is 0.7425 which is a good fit. In other words 74.25% of the dependent variable(y-values that means Balance) are explained by the independent variable(x-values that means credit limit).

3) Significance of F is the P-value of F.F is the F statistics or F-test for the null hypothesis .It is very effective used to test the overall model Significance.

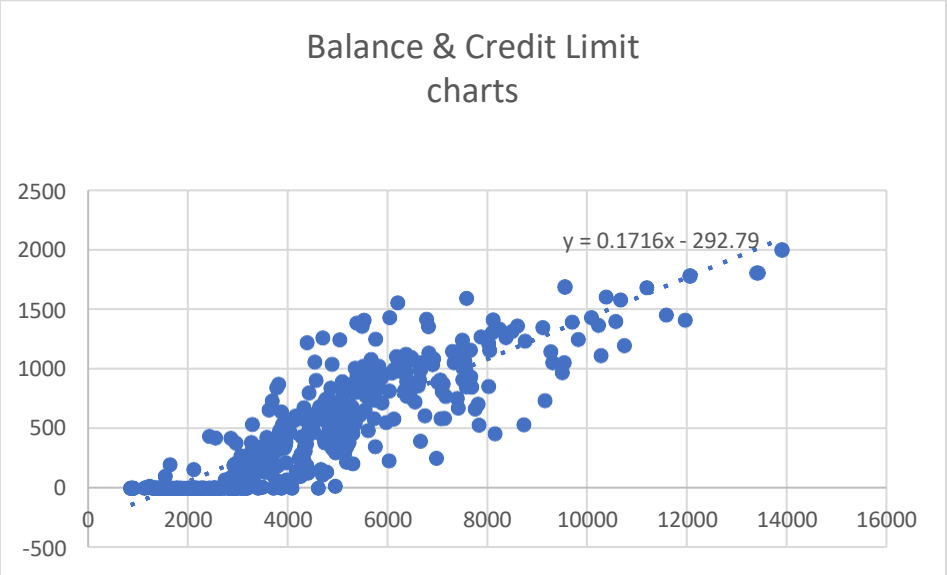
Regression Statistics	
Multiple R	0.861697267
R Square	0.74252218
Adjusted R Square	0.741875251
Standard Error	233.5849982
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6262425.25	6.3E+07	1147.764	2.5306E-119
Residual	398	21715656.66	54562		
Total	399	84339911.91			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-292.7904955	26.68341452	-10.9728	1.18E-24	-345.2485494	-240.3324415	-345.2485494	-240.3324415
credit_Limit	0.171637278	0.005066234	33.8787	2.5E-119	0.161677354	0.181597203	0.161677354	0.181597203

Data Science Individual Assessment



	credit Limit	*Balance*
credit Limit	1	
Balance	0.861697267	1

Statistics Point of View for Scatterplot Chart.

1) co-relation is a number between -1 and +1.
2) "+1" indicates a perfect strong positive linear relationship between X and Y.
3) "-1" indicates a perfect strong negative linear relationship between X and Y.
4) 0 indicates either no relationship or a complex nonlinear relationship.
5) The co-efficient value(0.861697267) for balance is towards the positive 1.
6) There is **strong positive relationship** between credit Limit & Balance.

*Calculation part in tab (answer-8(Limit&Balance)) in excel sheet.

9) Run a simple linear regression of balance (Y) on credit rating (X). Report the coefficients and R-squared. Show a scatter plot?

Ans:- we got the following table of results from this analysis.
Statistics Point of View:- (y= a+bx)

1) P- value is less than 0.05 so coefficient is significantly different from zero.
2) R-squared signifies the coefficient of determination which shows the goodness of fit.it shows how many points fall on the regression line. In our case the value of R-square is 0.74584 which is a good fit. In other words 74.58% of the dependent variable(y-values that means Balance) are explained by the independent variable(x-values that means credit Ratings).
3) Significance of F is the P-value of F.F is the F statistics or F-test for the null hypothesis .It is very effective used to test the overall model Significance.

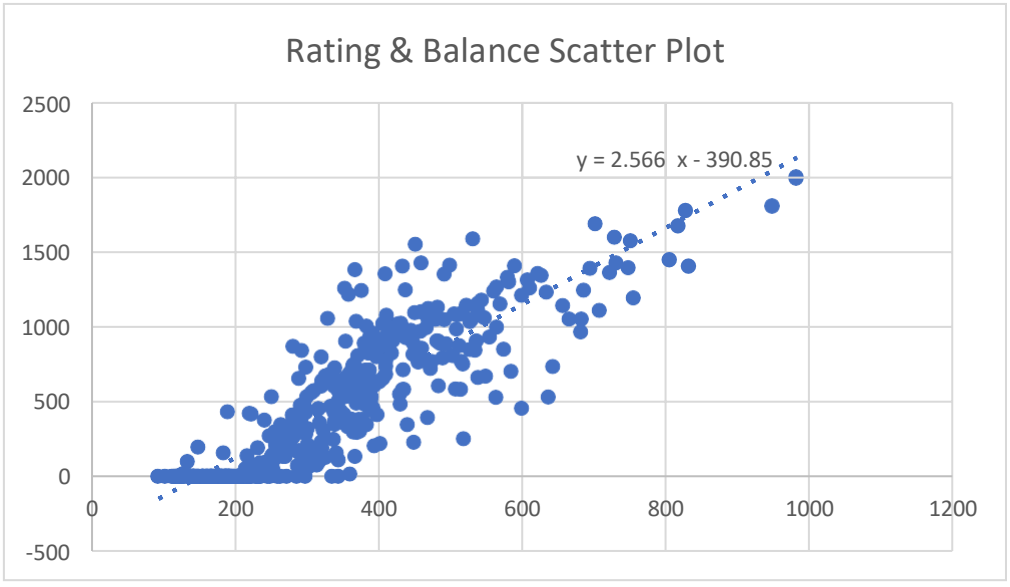
Data Science Individual Assessment

Regression Statistics	
Multiple R	0.863625161
R Square	0.745848418
Adjusted R Square	0.745209846
Standard Error	232.0713048
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62904789.88	62904789.88	1167.994581	1.8989E-120
Residual	398	21435122.03	53857.09053		
Total	399	84339911.91			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-390.8463418	29.06851463	-13.44569362	3.07E-34	-447.993365	-333.6993186	-447.993365	-333.6993186
Rating	2.566240327	0.075089102	34.17593357	1.8989E-120	2.418619483	2.713861171	2.418619483	2.713861171



	Rating	*Balance*

Data Science Individual Assessment

Rating	1	
Balance	0.863625161	1

Statistics Point of View for Scatterplot Chart.

1) co-relation is a number between -1 and +1.
2) "+1" indicates a perfect strong positive linear relationship between X and Y.
3) "-1" indicates a perfect strong negative linear relationship between X and Y.
4) 0 indicates either no relationship or a complex nonlinear relationship.
5) The co-efficient value(0.863625161) for balance is towards the positive 1.
6) There is **strong positive relationship** between Ratings & Balance.

*Calculation part in tab (answer-9(rating&Balance) in excel sheet.

10) Consider your findings in questions 8-9. Discuss business mechanisms to increase or decrease the balance on credit cards. Try to quantify your answers. In this context, focus on possible specific strategies using variables in Q8 and Q9 that the business could adopt to increase the balance on credit cards

Ans:-From Q8 and Q9 we can say in the following ways:-

1) The calculation is done based on the ratings & limit
2) If the credit limit increase one dollar then the average balance will increase by 0.171.
3) the credit rating is increase by one dollar then the average balance is increased by 2.566
4) Increase the credit rating of the customer, so that they can see the balance on credit cards.
5) The Multiple R is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger is the relationship. In Q8 Multiple R value is 0.861697267015396 that means Credit Limit & Balance has a positive stronger relationship where as in Q9 Multiple R value is 0.863625160621496 that means Ratings & Balance also a positive stronger relationship.so we can increase Limit & rating as well for increasing balance on cards.

6) in Q8 the value of R-square is 0.742522179981802 which is excellent fit with Limit & Balance. In Q9 the value of R-square is 0.745848418058504 which is also excellent fit with Ratings & Balance. Increase the credit Limit, increase the Balance from Q8 and increase the credit ratings also increase the balance.

*Calculation part in tab (answer-10) in excel sheet.

11 **The credit limit is provided as a consolidated amount for all the credit cards the cardholder has. Run a multiple linear regression of Balance (Y) on Limit and Cards as two X variables. Report the coefficients. Discuss the effect on the**

balance of (a) increasing the credit limit on the same number of cards and (b) increasing the number of cards without altering the total credit limit?

Ans: From the above analysis we got the following results.

Statistics Point of View :-

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.865188295
R Square	0.748550786
Adjusted R Square	0.747284 04
Standard Error	231.1247525
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	63132707.37	31566353.68	590.9238	9.7585E-120
Residual	397	21207204.54	53418.65124		
Total	399	84339911.91			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-369.035955 4	36.16414657	-10.20447018	7.23E-22	-440.133128	-297.939	-440.133	-297.939
Limit	0.171479037	0.005013136	34.20593861	2E-120	0.161623424	0.181335	0.161623	0.181335
Cards	26.03375427	8.438363509	3.085166246	0.002177	9.444290848	42.62322	9.444291	42.62322

1) R-squared signifies the coefficient of determination which shows the goodness of fit.it shows how many points fall on the regression line. In our case the value of R-square is 0.748550786206819 which is a good fit. In other words 74.85 or 75% of the dependent variable(y-values that means Balance) are explained by the independent variable(x-values that means credit limit & Cards). The p-value is less than 0.05, so we should reject the null hypothesis in favour of alternative

hypothesis. The population corelation coefficient is significantly different from zero in case of alternative hypothesis.

2) if we are increasing the credit limit the one dollar then the number of cards as constant then the average balance limit will be increased by 0.17.

3) if we are increasing the cards the one dollar then the limit as constant then the average balance cards will be increased by 26.033.

4) Null Hypothesis H0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.

5) Alternate Hypothesis Ha: The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

*Calculation part in tab (answer-11-multiple-linear-model) in excel sheet.

12) Run a simple linear regression equation with Income as X and Balance as Y. Report the coefficients. Is the coefficient of Income significantly different from zero? What does this say about the effect of income on balance?

Ans:- Based on the above linear regression we got the following model information

Regression Statistics	
Multiple R	0.463656457
R Square	0.214977031
Adjusted R Square	0.213004891
Standard Error	407.8647195
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18131167.4	18131167.4	108.9917	1.03E-22
Residual	398	66208744.51	166353.6294		
Total	399	84339911.91			

Data Science Individual Assessment

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	246.5147506	33.19934735	7.425289058	6.9E-13	181.2467	311.7828	181.2467	311.7828
Income	6.048363409	0.579350163	10.43990973	1.03E-22	4.909394	7.187332	4.909394	7.187332

1) **Report the Co-efficient**:-

The p-value is less than 0.05, so we should reject the null hypothesis in favour of the alternative hypothesis.

a) R Square signifies the Coefficient of Determination, which shows the goodness of fit. It shows how many points fall on the regression line. In our case , the value of R square is 0.21, which is not excellent fit between income & balance. We can't established strong relationship by the R-Square.

2) Is the coefficient of Income significantly different from zero ?

Null Hypothesis H0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.

Alternate Hypothesis H1: The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

In this case we p-value (1.03E-22) less than 0.05. Reject H0 in favour of H1. Enough evidence to say that the coefficient of income significantly different from zero.

3) What does this say about the effect of income on balance?

a) The Multiple R is the Correlation Coefficient that measures the strength of a linear relationship between income & Balance

The larger the absolute value, the stronger is the relationship. In this case Multiple R (0.463656457015758) has not established strong relationship between income & balance

b) R Square signifies the Coefficient of Determination, which shows the goodness of fit. It shows how many points fall on the regression line. In our case , the value of R square is 0.21, which is not excellent fit between income & balance. We can't established strong relationship by the R-Square.

*Calculation part in tab (answer-12(income&Balance)) in excel sheet.

Data Science Individual Assessment

13) Based on the equation derived in question 12, what is the estimated balance for a person with an income of USD 100k per year?

the estimated balance for a person with an income of USD 100k per year is **851.3510914**, the formula screenshot is given below : -

E	F	G	H	I	J	K	L	M
Adjusted R Square	0.213004891							
Standard Error	407.8647195							
Observations	400							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	18131167.4	18131167.4	108.9917152	1.03089E-22			
Residual	398	66208744.51	166353.6294					
Total	399	84339911.91						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	246.5147506	33.19934735	7.425289058	6.90E-13	181.2467485	311.7827527	181.2467485	311.7827527
Income	6.048363409	0.579350163	10.43990973	1.03089E-22	4.909394402	7.187332415	4.909394402	7.187332415
The estimated balance is								
=F18*100+F17								

*Calculation part in tab (answer-13(estimated balance)) in excel sheet.

14) Based on the dataset, explore the relationship between credit card balance (Y) and (a) Income (b) Age (c) Education (c) Limit, and (d) Rating as X variables? Estimate a multiple linear regression model and report the statistical significance of each of these variables.

Ans:- Based on the above multiple linear regression we got the following model information.

Regression Statistics	
Multiple R	0.936702578
R Square	0.877411720
Adjusted R Square	0.875856031
Standard Error	161.9917647
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*

Data Science Individual Assessment

Regression	5	74000827.17	14800165	564.0020686	4.5908E-177
Residual	394	10339084.74	26241.33		
Total	399	84339911.91			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-473.2514026	55.10833546	-8.58766	2.08837E-16	-581.5945666	-364.9082387	-581.5945666	-364.9082387
Income	-7.608832003	0.38193 1562	-19.922	1.37077E-61	-8.359710677	-6.857953333	-8.359710677	-6.857953333
Age	-0.860030445	0.47870 0493	-1.7965 9	0.07316 5937	-1.801157147	0.08109 6257	-1.801157147	0.08109 6257
Education	-1.967791521	2.60529 0902	-0.7553 06	0.45051 6748	-7.089801776	3.15421 8733	-7.089801776	3.15421 8733
Limit	-0.079016421	0.04479 1005	-1.7641 14	0.07848 7737	-0.167075679	0.00904 2839	-0.167075679	0.00904 2839
Rating	2.773843725	0.66707 9559	4.1581 9	3.93909E-05	1.462363177	4.085324273	1.462363177	4.085324273

In our regression model we can say that

A) The Multiple R is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger is the relationship.

B) R Square signifies the Coefficient of Determination, which shows the goodness of fit. It shows how many points fall on the regression line. In our example, the value of R square is 0.93, which is an excellent fit. In other words, 93% of the dependent variables (y-values) are explained by the independent variables (x-values).
F is the F statistic or F-test for the null hypothesis. It is very effectively used to test the overall model significance. +1 means strong positive relationship, whereas -1 means strong negative relationship & 0 means no relationship at all.

C) The Multiple R is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger is the relationship.

D) Adjusted R Square is the modified version of R square that adjusts for predictors that are not significant to the regression model.

E) ANOVA stands for Analysis of Variance. It gives information about the levels of variability within your regression model. Df(5) is the number of degrees of freedom associated with the sources of variance. SS is the sum of squares. The smaller the Residual SS viz a viz the Total SS, the better the fitment of your model with the data.

MS is the mean square. F is the F statistic or F-test for the null hypothesis. It is very effectively used to test the overall model significance. Significance F is the P-value of F.

The p value (1.37077E-61) of income (x axis with independent variable) with Balance (y axis with dependent variable) is less than 0.05, by this we can reject h0 or null hypothesis in favour of alternative hypothesis. Enough evidence to say that the population average balance depends on income, means it impact on balance.

The p value (0.0731659366159233) of Age (x axis with independent variable) with Balance (y axis with dependent variable) is greater than 0.05, by this we fail to reject the null hypothesis in favour of alternative hypothesis. We have not enough evidence to say that the population average balance is depends on Age. So, it does not create any impact on balance.

The p value (0.450516748443129) of Education (x axis with independent variable) with Balance (y axis with dependent variable) is greater than 0.05, by this we fail to reject the null hypothesis in favour of alternative hypothesis. We have not enough evidence to say that the population average balance is depends on Education. So, it does not create any impact on balance.

The p value (0.078487737397302) of Limit (x axis with independent variable) with Balance (y axis with dependent variable) is greater than 0.05, by this we fail to reject the null hypothesis in favour of alternative hypothesis. We have not enough evidence to say that the population average balance is depends on Limit. So, it does not create any impact on balance.

The p value (0.0000393909070412275) of Rating (x axis with independent variable) with Balance (y axis with dependent variable) is less than 0.05, by this we can reject h0 or null hypothesis in favour of alternative hypothesis. Enough evidence to say that the population average balance depends on Rating, means it impact on balance.

So finally, we can say that The p value of income & ratings is less than 0.05, by this we can say they have impact on balance.

*Calculation part in tab (Answer-14(multiLinearModel)) in excel sheet.

Data Science Individual Assessment