Q1.

Summary of Income Data:

50 persons are included in the "Income ($1000s)" data set as a sample. The sample's mean income of $43,48,000 indicates that the sample's average yearly income is around $43,480. The median income for the sample is $42,000, which indicates that half of the people earn less than this amount and half earn more. The sample's median income, or the most prevalent amount, is $54,000.

The standard deviation of the sample is $14.55k, which shows that there is a considerable amount of variance in the participants' wages. With the lowest income being $21,000 and the highest being $67,000, the range, or gap between the maximum and minimum incomes, is $46,000.

The data's skewness is positive but close to zero at 0.096, thus the distribution is somewhat to the right but is still relatively symmetrical. Since the kurtosis is negative at -1.248, the distribution seems to have thinner tails and fewer peaks than a normal distribution. The sample variance is 211.72, which expresses how scattered the data are around the mean.

Summary of Household Size Data:

The sample consists of 50 homes. There are often between 3 and 4 individuals living there because the sample's average household size is 3.42 persons. Half of the sample's residences have three or less occupants, while the other half have four or more, based on the median household size of three. The sample's median, or most average, household size is two individuals.

The sample's standard deviation is 1.74, thus there could be some variance in the sample's composition of households. The size of a home can range from 1 to 7, with the smallest and largest households having a 6 person difference.

According to the data's positive skewness of 0.53, the distribution is somewhat skewed to the right, which implies that there are more families with larger sizes. The negative kurtosis of -0.72 indicates that the distribution is less peaked and has thinner tails than a normal distribution. The sample variance, which describes the degree of data dispersion with respect to the average dwelling size, is 3.02.

Summary of Amount Charged Data:

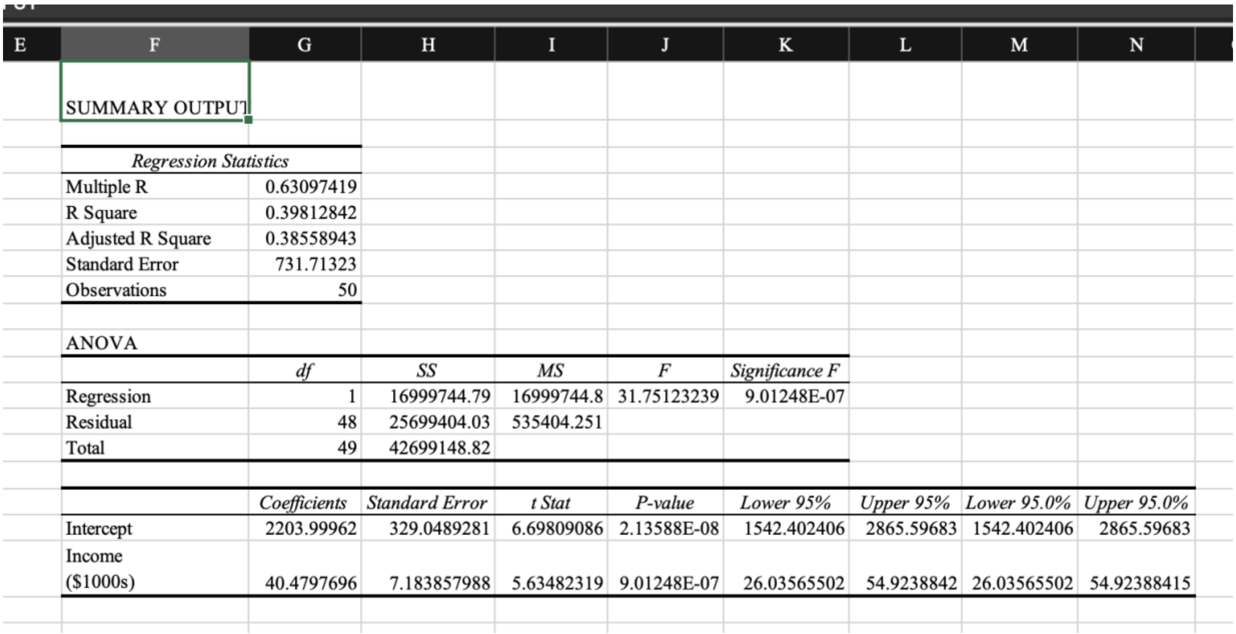
The sample "Amount Charged ($)" consists of 50 transactions. The mean amount charged indicates that the usual transaction in the sample had a value of around $3964.06. The median amount charged is a little bit higher at $4090, indicating that half of the transactions in the sample had a smaller amount and half had a bigger amount. The mode, or average, charge in the sample was $3890.

The sample's $933.49 standard deviation shows that prices charged for transactions in the sample vary significantly from one transaction to the next. There is a $3814 gap between the lowest and highest amounts charged, which range from $1864 to $5678.

According to the data's negative skewness of -0.13, there are more transactions with bigger amounts charged since the distribution is somewhat skewed to the left. The negative kurtosis of -0.74 indicates that the distribution looks to have thinner tails and a lower peak than a normal distribution. 871,411.20 is the sample variance, which measures how evenly distributed the data is around the mean amount charged.

Q2.

The regression equation can be written as Y = 2203.999618 + 40.47976959 * X1, where X1 stands for annual income and Y for the dependent variable.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.63097419
R Square	0.39812842
Adjusted R Square	0.38558943
Standard Error	731.71323
Observations	50

ANOVA

	df	SS	MS	F	Significance F
Regression	1	16999744.79	16999744.8	31.75123239	9.01248E-07
Residual	48	25699404.03	535404.251		
Total	49	42699148.82			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2203.99962	329.0489281	6.69809086	2.13588E-08	1542.402406	2865.59683	1542.402406	2865.59683
Income ($1000s)	40.4797696	7.183857988	5.63482319	9.01248E-07	26.03565502	54.9238842	26.03565502	54.92388415

Figure a: Detailed Statistics for Income

The multiple R value of 0.6309 indicates a moderately strong positive linear relationship between the dependent variable and annual income. The change in yearly income accounts for 39.8% of the variance in the dependent variable, according to the R-squared value of 0.3981.

The coefficients table provides specifics on the regression line's intercept and slope. The intercept, which equals 2203.999618, or 2203 for short, represents the value of the dependant variable when the income is zero. Although the minimal income value for the dataset is 21, the intercept has no practical relevance. The slope coefficient is 40.47976959, which means that for every $1 rise in annual income, the dependent variable increases by $40.48.

The ANOVA table indicates that the regression model is significant, with an F- statistic of 31.751 and a corresponding p-value of 9.01248E-07. This demonstrates that the regression model significantly improves the prediction of the dependent variable when compared to a model with no independent variables.

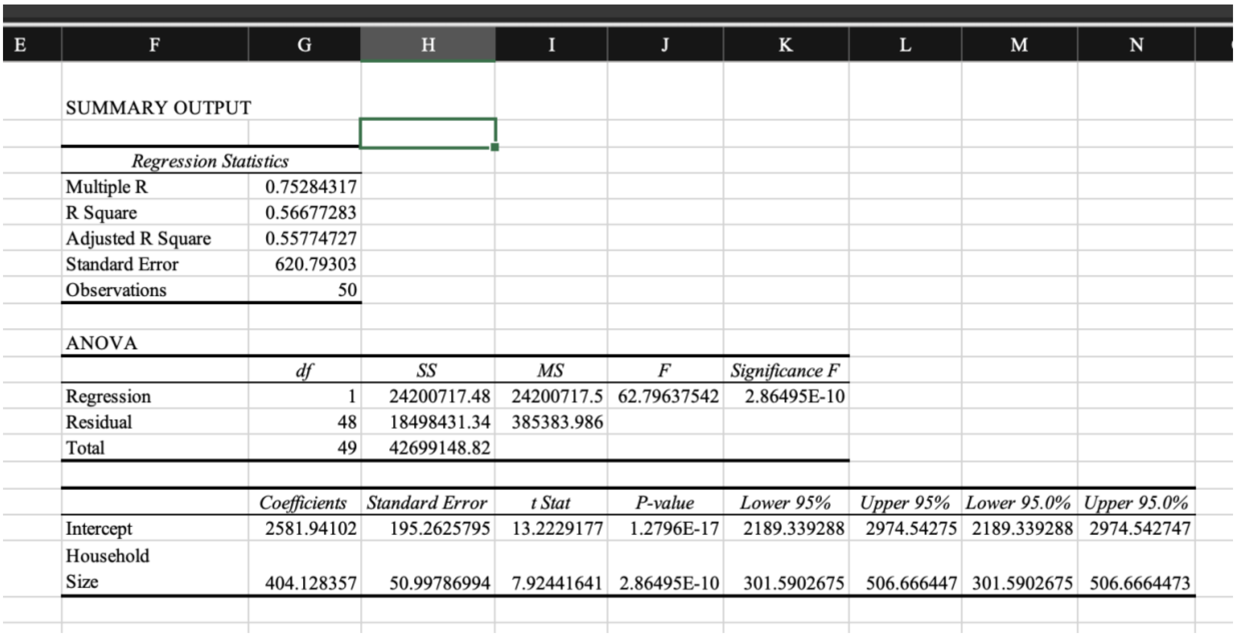
According to the calculated regression equation, annual income strongly predicts the dependent variable, with higher income being connected with higher values of the dependent variable. It is important to keep in mind that the regression model does not necessarily imply causality between the variables and that other factors may potentially have an effect on the dependent variable.

Q3.

The derived regression equation, where Y is the dependent variable and X2 represents household size, is provided by:

Y = 2581.94 + 404.13 * X2.

Household size is denoted by X2, and X2 acts as the independent variable.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.75284317
R Square	0.56677283
Adjusted R Square	0.55774727
Standard Error	620.79303
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	24200717.48	24200717.5	62.79637542	2.86495E-10
Residual	48	18498431.34	385383.986		
Total	49	42699148.82			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2581.94102	195.2625795	13.2229177	1.2796E-17	2189.339288	2974.54275	2189.339288	2974.542747
Household Size	404.128357	50.99786994	7.92441641	2.86495E-10	301.5902675	506.666447	301.5902675	506.6664473

Figure b: Detailed Statistics for Household Size

This shows that for every additional unit of household size, the predicted yearly income increases by $404.13, assuming all other factors stay constant.

The R-squared value of 0.57 indicates that the variance in annual income is 57% due to the difference in family size. As a consequence, although it is likely that other factors also have an influence, it appears that household size is a significant predictor of annual income.

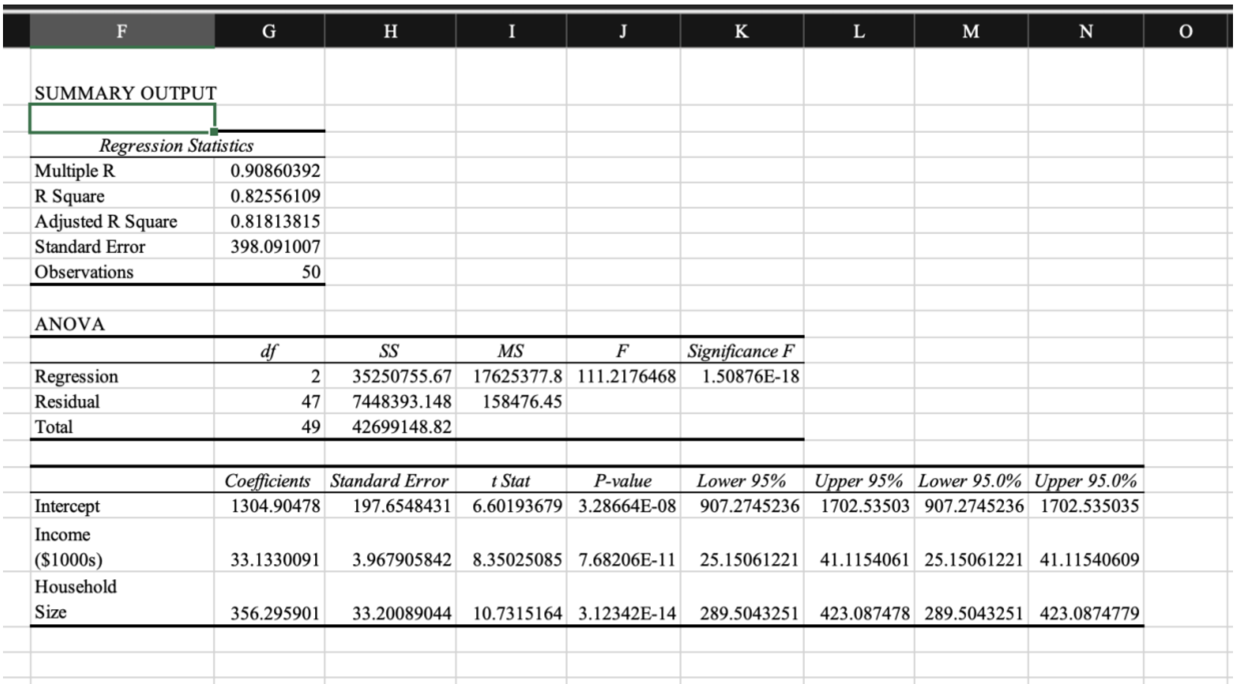
The regression model's statistical significance and the fact that at least one of the independent variables strongly predicts the dependent variable are both shown by the F-test's extraordinarily low p-value (2.86E-10).

The Household Size Coefficient has a p-value of 2.86E-10 and a t-statistic of 7.92. The fact that the p-value is significantly less than the typical alpha level of 0.05 allows us to draw the conclusion that household size significantly predicts annual income. Based on the 95% confidence interval for the Household Size coefficient, which is (301.59, 506.67), we may be 95% confident that the true effect of household size on income falls within this range.

In general, the findings show that household size is a significant predictor of yearly income and that the generated regression equation may accurately predict income based on household size. Given that the model only accounts for 57% of the variation in income, it's critical to keep in mind that additional factors could also have an impact.

Q4.

The calculated regression equation, which takes yearly income and household size into account as independent variables, is as follows: Y = 1304.904779 + 33.13300915X1 + 356.2959015X2, where Y stands for the dependent variable (total annual expenditures), X1 for annual income, and X2 for household size.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.90860392
R Square	0.82556109
Adjusted R Square	0.81813815
Standard Error	398.091007
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	35250755.67	17625377.8	111.2176468	1.50876E-18
Residual	47	7448393.148	158476.45		
Total	49	42699148.82			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1304.90478	197.6548431	6.60193679	3.28664E-08	907.2745236	1702.53503	907.2745236	1702.535035
Income ($1000s)	33.1330091	3.967905842	8.35025085	7.68206E-11	25.15061221	41.1154061	25.15061221	41.11540609
Household Size	356.295901	33.20089044	10.7315164	3.12342E-14	289.5043251	423.087478	289.5043251	423.0874779

Figure c: Detailed Statistics for Income & Household size

The multiple R value of 0.9086 indicates a substantial positive correlation between the independent variables and the dependent variable. With an R-squared value of 0.8256, the linear relationship between the independent factors and the dependent variable explains almost 83% of the variation in total annual expenditures.

The F- score of 111.2176 and the p-value of 1.50876E-18 in the ANOVA table show that the regression is statistically significant. This demonstrates that the independent variables and the dependent variable have a significant positive connection.

By looking at the coefficients, we can see that both the annual income and the size of the family have a positive and statistically significant impact on the total annual expenditures. For every $1,000 increase in income and for each new member of the household, total annual costs rise by $33.13 and $356.30, respectively.

In the end, the calculated regression equation reveals a strong positive association between annual income, household size, and total annual spending. The model can explain a sizeable fraction of the variation in total annual expenditures and both independent variables play a significant role in predicting total annual expenditures.

Q5.

To determine which regression equation is best at predicting annual credit card charges, we must consider a number of factors.

Since it shows how much of the variance in the dependent variable (credit card charges) can be attributable to the independent variables (income and family size), the R-squared value is a good place to start. The R-squared number increases with the accuracy of the regression line's fit to the data.

Regression 1/ Question 2, Regression 2/ Question 3, and Regression 3/ Question 4 had respective R-squared values of 0.398, 0.567, and 0.826. Because Regression 3 has the highest R-squared value, we may infer that it best explains the variance in credit card charges.

The second thing we may look at is the significance level of the regression coefficients. The p-value can be used to assess the statistical significance of the regression coefficients. It is considered significant if a coefficient's p-value is less than 0.05.

In Regression 1, the coefficient of income is significant but not the intercept, whereas in Regression 2, the coefficient of household size is significant but not the intercept.

Conversely, the coefficients for the income, household size, and intercept in Regression 3 are all statistically significant. Family size and income are both highly significant predictors of credit card charges, therefore it stands to reason that Regression 3 gives the best match for doing so.

We may thus deduce that Regression 3 is the most effective at forecasting yearly credit card costs based on both the R-squared value and the importance of the coefficients.

Q6.

We may use Regression 3 to estimate the annual credit card cost by entering the data for family size and income. In this case, there are 3 family members, and the family's yearly income is $40,000. Consequently, we have

(Rounded to the next two decimal places) Predicted Annual Credit Card Charge 1304.90 + 33.13 * 40 + 356.29 * 3 = 1304.90 + 1325.32 + 1068.88 = $3699.11

Therefore, the expected annual credit card fee for a three-person household making $40,000 is $3699.11.