HULT
INTERNATIONAL
BUSINESS SCHOOL

Team Project

Final Report

MSBA Fall 2020

Advanced Data Analytics

Team 1

YUYANG CAI

Patrick Kincaid

Giulia Mosiewicz

Valentin Voelckel

Fabiola Farrera Fonseca

Cleaning the dataset & defining assumptions

Outliers:

The number of observations in the data set was adjusted from 2241 to 2237 by deleting four Outliers:

- **Year_Birth**: Three rows were deleted with observations indicating the customer was born before 1900 (Invalid Customer data is being assumed)
- **Income:** One row was deleted, indicating income of 666666, which is over four times bigger than the second next highest value.

Blanks:

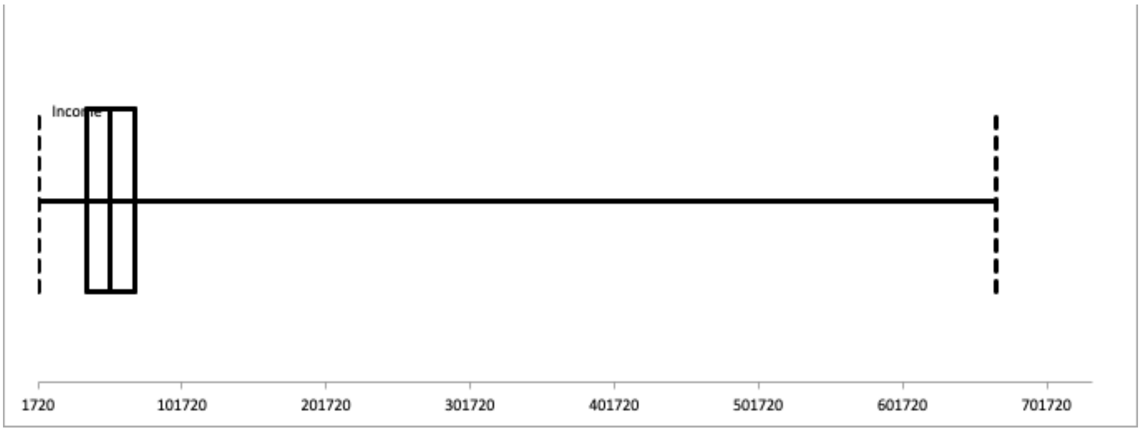
From 24 to 0 blanks:

- Assumption: The data provided by customers was correct, but income was too sensitive information for customers to provide. For the 24 missing, values the **mean** per education was imputed for affected cells.

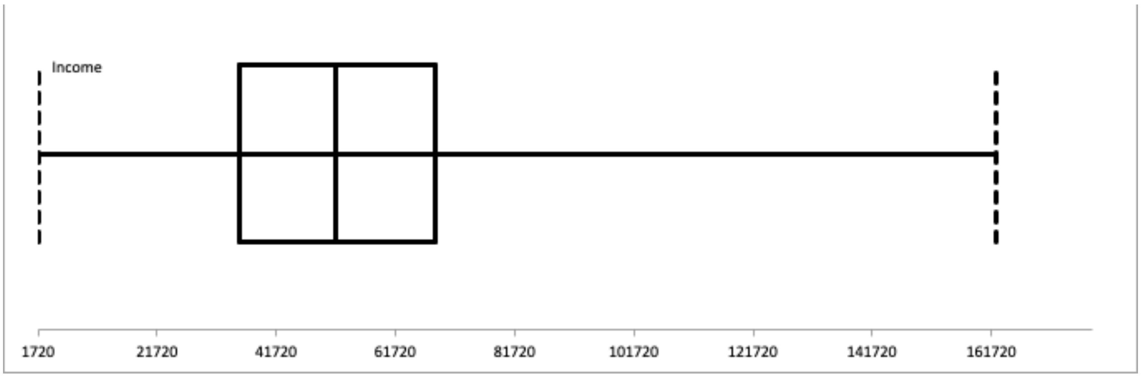
Income:

Income is the most affected variable from cleaning the data set. The impact on the distribution of the data by deleting the outlier and imputing values is visualized in the following:

Before:



After:



After imputing the missing values, the distribution of the income variable is now closer to a normal distribution, which ensures more accurate insights.

Dummy Variables:

Name Dummy Var	**Based on**	**Reason/use**
Age: Transform Date of birth into age Assumption we are in Year 2014.	Year_Birth	Simplify calculations
DaysCustomer: Count days for how long a customer exists until 2014/07/01	Dt_Customer	See if long term customers behave differently in behavior
SumAcceptetOffer	AcceptedCmp 1-5	See if offers were overall effective
IncDummy	Income	Check how income groups compare

Data set overview

Columns:

- Z_Revenue Z_CostContact were deleted as they were not defined and redundant.
- The first three columns ID, Outlier Income and Dt_Customer are only used for general analysis but are not used for regression.
- Color coding in the Excel sheet DataCleaned categorizes variables according to the overview explained in sheet VarDef.

Part I: Statistical tests and Regression

A1) What factors are significantly related to the number of web purchases?

To evaluate significantly related factors for web purchases, we test how different factors correlate with the responding variable. The selection for our multiple regression model is based on the logic of what variables might influence the online purchases as well as a chain of the following analysis. Computing a correlation matrix (ACorrelationMatrix) gives us an overview of relating factors. In the matrix, It can be seen that the variables with the highest correlation to web purchases are the amount of wine purchased (0.54), followed by income (0.45), amount of gold purchased (0.42), and a negative relationship for the number of kids in a household with -0.36. Consequently, these and other variables are tested with linear regression and visually inspected with scatterplots. For example, in this step, we define income as a significant predictor for web purchases, given a significance level of 2.35E-112 (Graph 1). Despite a few outliers in income at around 160.000, we observe that higher income leads to greater purchase power on the internet.

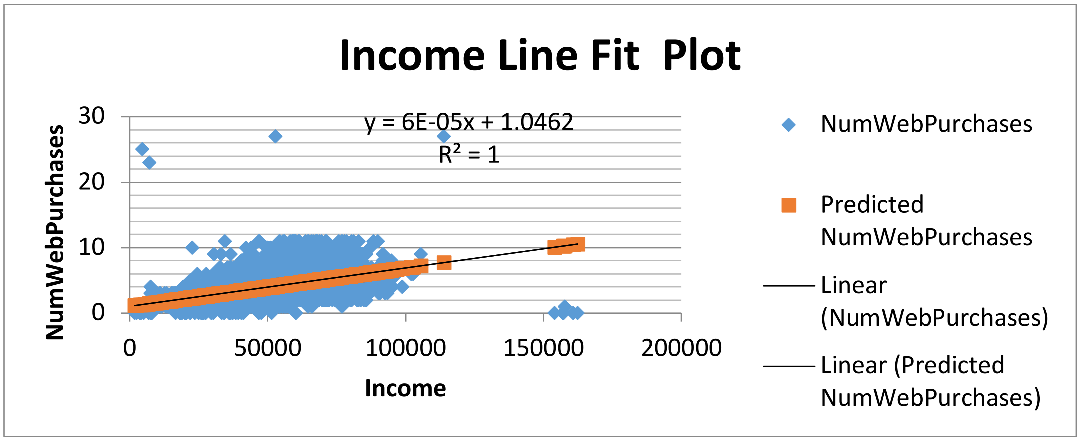
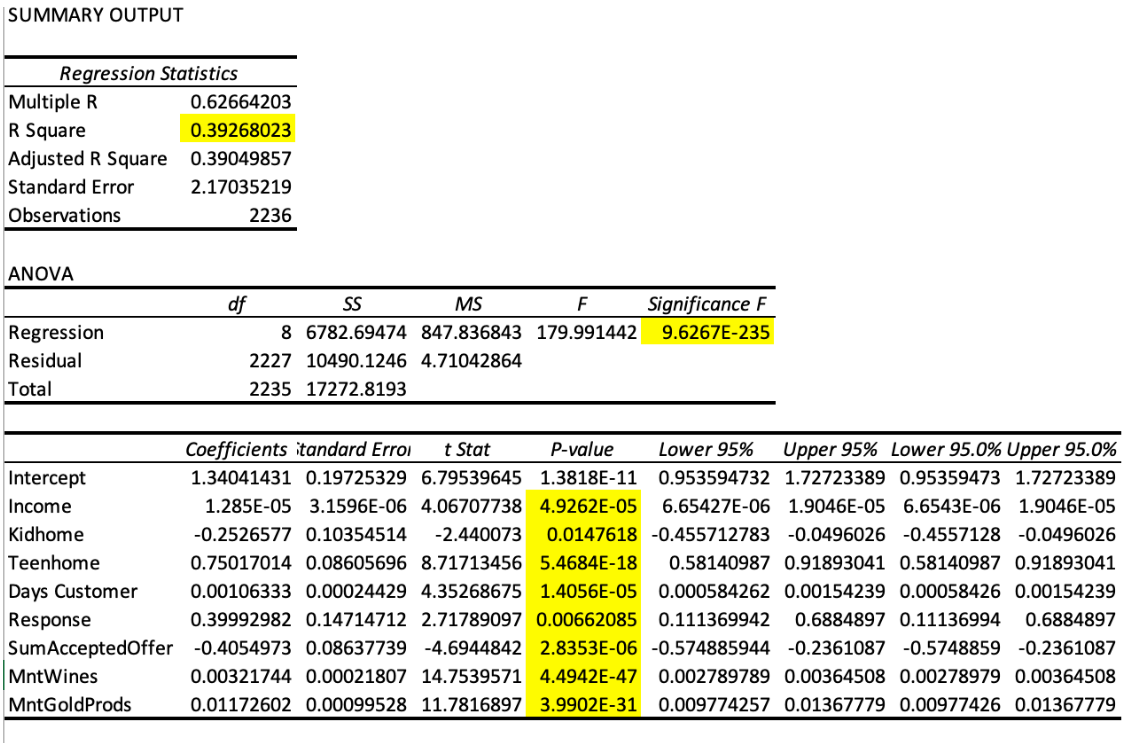


Chart 1: Income Line Fit Plot

After conducting linear regressions, we use significant findings to build a multiple regression model, presented in the following.

Table 1:Output Multiple Regression Web-Purchases



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.62664203
R Square	0.39268023
Adjusted R Square	0.39049857
Standard Error	2.17035219
Observations	2236

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	6782.69474	847.836843	179.991442	9.6267E-235
Residual	2227	10490.1246	4.71042864		
Total	2235	17272.8193			

	Coefficients	*tandard Erroı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.34041431	0.19725329	6.79539645	1.3818E-11	0.953594732	1.72723389	0.95359473	1.72723389
Income	1.285E-05	3.1596E-06	4.06707738	4.9262E-05	6.65427E-06	1.9046E-05	6.6543E-06	1.9046E-05
Kidhome	-0.2526577	0.10354514	-2.440073	0.0147618	-0.455712783	-0.0496026	-0.4557128	-0.0496026
Teenhome	0.75017014	0.08605696	8.71713456	5.4684E-18	0.58140987	0.91893041	0.58140987	0.91893041
Days Customer	0.00106333	0.00024429	4.35268675	1.4056E-05	0.000584262	0.00154239	0.00058426	0.00154239
Response	0.39992982	0.14714712	2.71789097	0.00662085	0.111369942	0.6884897	0.11136994	0.6884897
SumAcceptedOffer	-0.4054973	0.08637739	-4.6944842	2.8353E-06	-0.574885944	-0.2361087	-0.5748859	-0.2361087
MntWines	0.00321744	0.00021807	14.7539571	4.4942E-47	0.002789789	0.00364508	0.00278979	0.00364508
MntGoldProds	0.01172602	0.00099528	11.7816897	3.9902E-31	0.009774257	0.01367779	0.00977426	0.01367779

The Regression (Table 1) model includes eight variables:

- Income
- Kidhome (Number of kids in the household)
- Teenhome (Number of teens in the household)
- Days Customer (Dummy for the term of customers)
- Response (Customer accepted in the last campaign)
- SumAccepted Offer (Dummy for the sum of accepted offers)
- MntWines (Amount of wine purchased)
- MntGoldProds (Amount of gold products purchased)

The R-Square value of 39% indicates that these variables predict over one-third of the dependent variable variation. Significance F score of 9.62E-235 and the individual P-values for each variable, which are lower than alfa 0.05 for the 95% confidence level. The Coefficients allow us to draw conclusions from a predictive standpoint, where the number of kids at home and the sum of accepted offers are the only two negative indicators for web-purchases. For example, each additional kid in a household will decrease the number of web-purchases by .025. In contrast, web purchases are predicted to increase by 0.75 for every additional teenager at home. We also notice that the variable Response for the last campaign has a relatively strong impact on web-purchases, which might have been the marketing’s intention.

Finally, the regression model adding the coefficients together can be defined as the following:

$$\boxed{\boldsymbol{Y = 1.340 - 1.285E\text{-}05\,\beta_1 + -0.252\,\beta_2 + 0.750\,\beta_3 + 0.001\,\beta_4 + 0.399\,\beta_5 + -0.405\,\beta_6 + 0.003\,\beta_7 + 0.001\,\beta_8}}$$

The model’s intercept Y, which is the mean of the dependent variable, shows that 1.340 purchases will be made through the company’s website if the other explanatory variables are zero.

A2) What would you suggest to CMO to improve web purchases?

Analyzing the data, it seems that the marketing campaigns mainly failed due to improper customer targeting. The data shows a vast audience spread out over several countries was covered in marking efforts, which means that the marketing has to take several cultures with different preferences. Therefore, the company should focus resources on defining a niche and the right channel to market specific products. Based on the result of the multiple regression, a customer person for web purchases can be defined. Taken the positive impact of income and the number of teenagers into account, the group with represented attributes should be in the CMO’s focus. Tailored marketing campaigns, advertising Gold and Wine products promise success, given the positive coefficients.

Furthermore, the sum of accepted offers is predicted to decrease online sales and should be redesigned. Recent efforts, shown by the variable Response should be considered in the new marketing concept, as the regression determines a positive impact on web purchases. According to the findings, another critical factor is customer maturity. It is evident that purchases on the web raise with every day the company retains a customer. Efforts to offer deals about the online store and present information to redirect clients to the website are likely to improve the online channel. Besides, future marketing campaigns should be addressed to the primary Spanish market, while additional research for remaining countries has to be advanced to explore more country-specific patterns.

Summarizing, we suggest the CMO shift towards a more customer-centric marketing approach, defining groups based on demographic criteria such as income and purchasing behavior.

B) Does US fare significantly better than RoW (Rest of the World) in terms of total purchases.

To evaluate which category has more total purchases, dummy variables for the US and the Rest of the World were computed. After applying filters to the dataset, separating the countries, the total purchase number for each variable was copied into a new worksheet for further analysis. The Excel sheet (B Output) shows generated variables, resulting in two different sized samples of 2026 (ROW) and 1473 (USA).

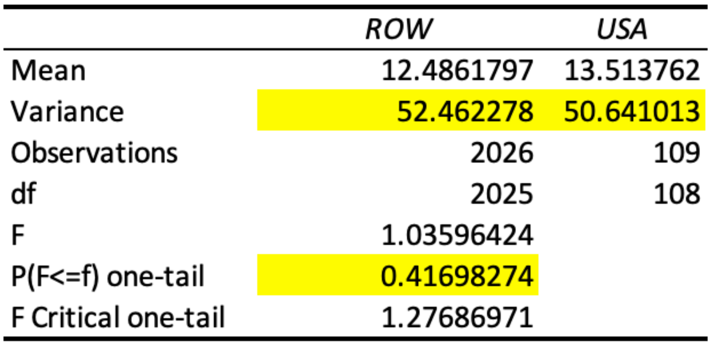
In order to decide whether the USA significantly fares better in total purchases, the following Hypothesis was created:

H0: μ USA - μ ROW <= 0

H1: μ USA - μ ROW >0

To compare the difference of means for two populations with unknown variances, an independent T-test was performed. By performing an F-Test for the two sample variances, illustrated in table 2, we detect similar variances (52.46 and 50.64). Additionally, the score for P(F<=f) proves that there is no significant difference in the population variances.

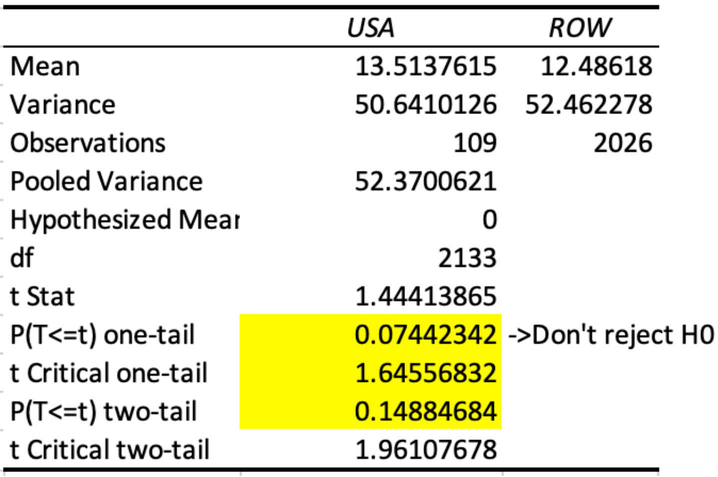
Table 2:F-Test for comparison ROW USA



	ROW	*USA*
Mean	12.4861797	13.513762
Variance	52.462278	50.641013
Observations	2026	109
df	2025	108
F	1.03596424	
P(F<=f) one-tail	0.41698274	
F Critical one-tail	1.27686971	

As a result, we conducted a T-test assuming equal variances, shown in the following.

Table 3:T-Test USA ROW



	USA	ROW	
Mean	13.5137615	12.48618	
Variance	50.6410126	52.462278	
Observations	109	2026	
Pooled Variance	52.3700621		
Hypothesized Mear	0		
df	2133		
t Stat	1.44413865		
P(T<=t) one-tail	0.07442342	->Don't reject H0	
t Critical one-tail	1.64556832		
P(T<=t) two-tail	0.14884684		
t Critical two-tail	1.96107678		

The computed results display a critical value of 1.645 for the one-tail test and a test statistic of 1.444. As the test statistic is not greater than the critical value, and the one-tailed P-value is greater than alfa 0.05, **we fail to reject H0**. We can conclude that there is no significant difference in the means of total purchases between the USA and the Rest of the World.

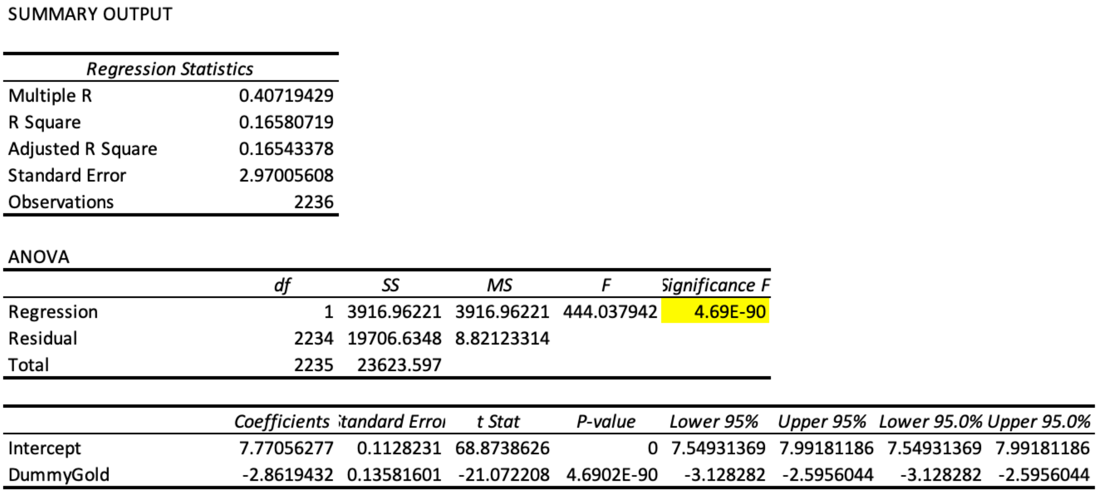
C) Your supervisor insists that people who buy gold are more conservative and as such people who spent an above average amount on gold in last 2 years would have more in store purchases. Justification/rebuttal:

H0: µ Gold Purchases > 44 do not increase In-Store Purchases

H1: µ Gold Purchases > 44 increase In-Store Purchases

Linear regression was performed to reject or fail to reject the hypothesis H0 that customers with an above-average of gold purchases do not increase in-store purchases. For this purpose, a dummy variable dividing the population into two groups with a mean above and below 44 was created and compared to the dependent variable (Table 4).

Table 4: Regression Output C Gold



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.40719429
R Square	0.16580719
Adjusted R Square	0.16543378
Standard Error	2.97005608
Observations	2236

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3916.96221	3916.96221	444.037942	4.69E-90
Residual	2234	19706.6348	8.82123314		
Total	2235	23623.597			

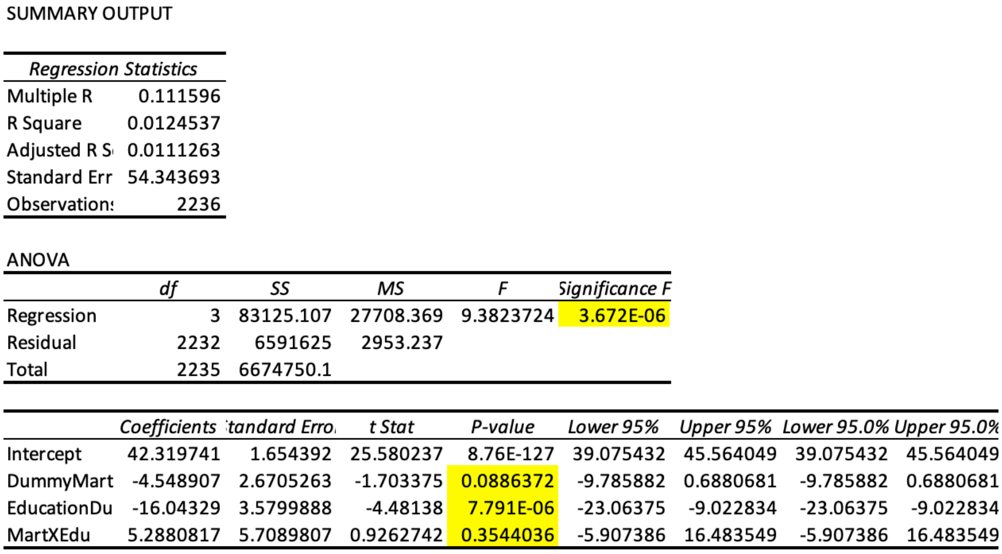
	Coefficients	tandard Erroı	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	7.77056277	0.1128231	68.8738626	0	7.54931369	7.99181186	7.54931369	7.99181186
DummyGold	-2.8619432	0.13581601	-21.072208	4.6902E-90	-3.128282	-2.5956044	-3.128282	-2.5956044

The Significance F score of 4.69E-90 indicates that there is a strong relationship between the two observed variables. Therefore, more gold purchases lead to higher in-store purchases, resulting in a rejection of H0. Therefore, the assumption that an increasing amount of gold purchased influences the store purchasing channel. However, we cannot say that this relationship is due to a more conservative behavior as assumed by the supervisor.

D) Fish has Omega 3 fatty acids, good for brain, accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish?

To examine if the amount spent on fish has a significant relation with the population's proportion of Married Ph.D. candidates, a multiple regression was conducted (Table 5).

Table 5: Regression Output Married Ph.D.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.111596
R Square	0.0124537
Adjusted R S	0.0111263
Standard Err	54.343693
Observation	2236

ANOVA

	df	SS	MS	F	Significance F
Regression	3	83125.107	27708.369	9.3823724	3.672E-06
Residual	2232	6591625	2953.237		
Total	2235	6674750.1			

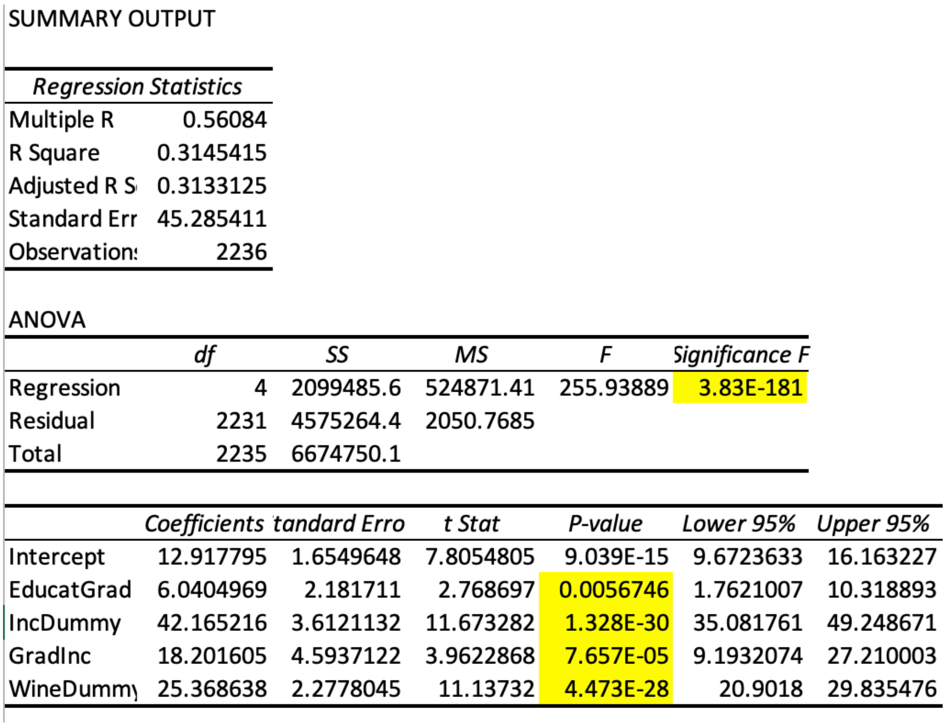
	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	42.319741	1.654392	25.580237	8.76E-127	39.075432	45.564049	39.075432	45.564049
DummyMart	-4.548907	2.6705263	-1.703375	0.0886372	-9.785882	0.6880681	-9.785882	0.6880681
EducationDu	-16.04329	3.5799888	-4.48138	7.791E-06	-23.06375	-9.022834	-23.06375	-9.022834
MartXEdu	5.2880817	5.7089807	0.9262742	0.3544036	-5.907386	16.483549	-5.907386	16.483549

Despite an overall significance of 3.672E-06 the multiple linear regression model **$Y = 42.319 + -4.548 \beta_1 + -16.043 \beta_2 + 5.288 \beta_3$** does not represent a significant association with the response variable Fish Purchases. The failed rejection of H0: $\beta_1 = \beta_2 = \beta_3 = 0$ is supported by individual non-significant p-values for the marital status and the multiplied explanatory variables. Furthermore, a low r-square that only around 12% of the selected variables the variation in the dependent variable. However, we observe a significant result for variable, Education Dummy representing Ph.D. candidates. To improve the regression model and identify possible confounding variables, further analysis is required to influence the given results.

D) What other factors are significantly related to amount spent on fish?

Based on the previous findings that education relates to fish purchases, we explore the higher education's impact with the most frequent education level Graduation. We also believe that income is significant within this context, as fish products are generally perceived as more expensive. Therefore, we integrate a dummy variable capturing households with an income above $70.000 into our new model and multiply it with the Graduation level in addition. Furthermore, a generated correlation matrix indicates that wine purchases affect the investigated response.

Table 6: Advance Multiple Regression Model



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.56084
R Square	0.3145415
Adjusted R S	0.3133125
Standard Err	45.285411
Observation:	2236

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2099485.6	524871.41	255.93889	3.83E-181
Residual	2231	4575264.4	2050.7685		
Total	2235	6674750.1			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.917795	1.6549648	7.8054805	9.039E-15	9.6723633	16.163227
EducatGrad	6.0404969	2.181711	2.768697	0.0056746	1.7621007	10.318893
IncDummy	42.165216	3.6121132	11.673282	1.328E-30	35.081761	49.248671
GradInc	18.201605	4.5937122	3.9622868	7.657E-05	9.1932074	27.210003
WineDumm)	25.368638	2.2778045	11.13732	4.473E-28	20.9018	29.835476

The regression output shows a substantial increase in the r-square to 0.3145 compared to the original model. Given that all factors are within the confidence level of 95% and their high coefficients, the selected variables explain a stronger relationship with the number of fish products being purchased.

E) Do any other analysis you deem relevant to show to your CMO. For the purpose, propose a hypothesis and perform the appropriate tests.

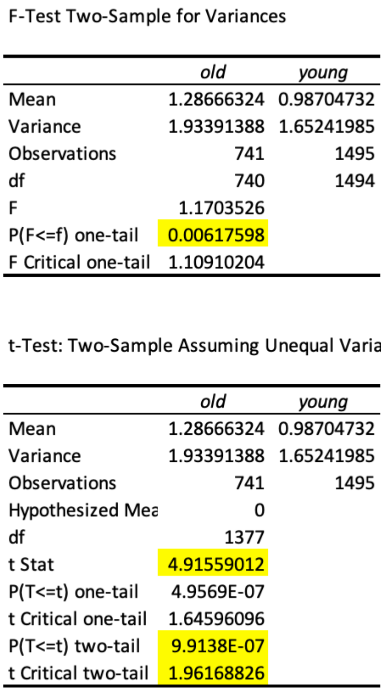
To get a deeper understanding of how age relates to online shopping behavior, we focus our final analysis on digital immigrants and digital natives. The latter group grew up with technology and is aged 50 and below. On the other hand, the Digital Immigrants had to adapt to technology, which might affect their ability to quickly navigate the company's website from the landing page to the check-out (dummy variable web-purchases/web-visits). Therefore, we hypothesize the following:

H0: µ younger - µ older = 0

H1: µ younger - µ older ≠ 0

To validate the hypothesis, we run a T-test for two sample means, comparing if defined age groups differ in their web-purchase to web-visit ratio (Table 7).

Table 7: F- T-Test Age Visits Web-Purchases



F-Test Two-Sample for Variances

	old	*young*
Mean	1.28666324	0.98704732
Variance	1.93391388	1.65241985
Observations	741	1495
df	740	1494
F	1.1703526	
P(F<=f) one-tail	0.00617598	
F Critical one-tail	1.10910204	

t-Test: Two-Sample Assuming Unequal Varia

	old	*young*
Mean	1.28666324	0.98704732
Variance	1.93391388	1.65241985
Observations	741	1495
Hypothesized Mea	0	
df	1377	
t Stat	4.91559012	
P(T<=t) one-tail	4.9569E-07	
t Critical one-tail	1.64596096	
P(T<=t) two-tail	9.9138E-07	
t Critical two-tail	1.96168826	

After the assessment to test for unequal variances (Table 1, F-Test) due to a significant difference in the sample means, the related T-test was executed (Table 2). Given the test statistic of 4.91, exceeding t critical two-tail value, **we reject H0**. There is enough evidence to conclude that digital immigrants need more web visits to purchase a product than digital natives. To improve the online visit to purchase ratio for customers aged above 50, the CMO is recommended to redesign the website more self-explaining for the older audience.