CMA:TEAM ASSIGNMENT
GROUP DS2

NYONGTAKUBAI ASHU AMPUE
ETERI KIKALISHVILI
TOLGA SUMER
MATTEO BONINI
SEBASTIAN ANUSCH

ANSWER #0

```python
# create a new column with AGE in ranges from 18-35, 36-65
df['Age_Range'] = pd.cut(df['Age'], bins=[18, 35, 65], labels=['18-35', '36-65'])
df.head()
```

✓ 0.0s

```python
# transform AGE_RANGE to dummies
df = pd.get_dummies(df, columns=['Age_Range'], drop_first=True)
df.head()
```

✓ 0.0s

```python
df['Utilitarian'] = np.where((df['Category'] == 2) | (df['Category'] == 0), 1, 0)
df['Hedonic'] = np.where((df['Category'] == 2) | (df['Category'] == 1), 1, 0)
df.head()
```

- CREATE AGE RANGES: 18-35 AND 36-65, TRANSFORM INTO DUMMY AND DROP AGE
- DROPPING GENERAL FEATURES FOR DESIGN,TECHNICAL,PRICE, SERV_DELIVERY, ZNUMBER_WORD_REVIEWS
- DIVIDING THE CATEGORY COLUMN INTO TWO DIFFERENT DUMMY COLUMNS CALLED “UTILITARIAN” AND “HEDONIC”

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Rating_Score   R-squared:                       0.685
Model:                            OLS   Adj. R-squared:                  0.676
Method:                 Least Squares   F-statistic:                     74.42
Date:                Wed, 20 Mar 2024   Prob (F-statistic):          1.56e-133
Time:                        21:35:32   Log-Likelihood:                -633.23
No. Observations:                 600   AIC:                             1302.
Df Residuals:                     582   BIC:                             1382.
Df Model:                          17
Covariance Type:            nonrobust
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6829	0.176	9.586	0.000	1.338	2.028
Number_Words_Review	-0.0004	0.001	-0.545	0.586	-0.002	0.001
Prod_Design_positive	0.1792	0.066	2.714	0.007	0.050	0.309
Prod_Design_negative	-0.0395	0.071	-0.556	0.578	-0.179	0.100
Prod_Technical_positive	0.7524	0.078	9.604	0.000	0.599	0.906
Prod_Technical_negative	-0.4577	0.082	-5.567	0.000	-0.619	-0.296
Prod_Price_positive	0.0135	0.076	0.177	0.859	-0.136	0.163
Prod_Price_negative	-0.0518	0.140	-0.369	0.712	-0.327	0.224
Serv_Delivery_positive	0.0468	0.153	0.305	0.760	-0.254	0.348
Serv_Delivery_negative	-0.4880	0.216	-2.263	0.024	-0.911	-0.065
Country	-0.1185	0.065	-1.832	0.067	-0.246	0.009
Gender	-0.0191	0.059	-0.323	0.747	-0.135	0.097
Sentiment	0.5014	0.035	14.386	0.000	0.433	0.570
Purchase	0.1339	0.113	1.187	0.236	-0.088	0.355
Number_of_Purchases	-0.0213	0.051	-0.418	0.676	-0.121	0.079
Age_Range_36-65	-0.0158	0.061	-0.260	0.795	-0.135	0.104
Utilitarian	0.1952	0.077	2.524	0.012	0.043	0.347
Hedonic	0.1987	0.075	2.632	0.009	0.050	0.347

```
==============================================================================
Omnibus:                        8.924   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.012   Jarque-Bera (JB):               10.166
Skew:                          -0.205   Prob(JB):                      0.00620
Kurtosis:                       3.489   Cond. No.                         467.
==============================================================================
```

ANSWER #1

- WE CHOOSE TO USE LINEAR REGRESSION BECAUSE THE DEPENDENT VARIABLE WHICH IS THE RATING SCORE IN OUR CASE IS CONTINUOUS AND CAN TAKE REAL VALUES FROM 1 TO 5.
- IDENTIFY “RATING SCORE” AS TARGET TO PERFORM LINEAR REGRESSION AND REMOVE IT + “REVIEW ID” FROM DATASET
- ANALYZE OLS REGRESSION RESULTS TO DETERMINE SIGNIFICANT PREDICTORS IN AFFECTING “RATING SCORE” --> P-VALUE <= 0.05, NAMELY:

1) PROD_TECHNICAL_POSITIVE --> 0.000 AND 0.7524 COEF
2)SENTIMENT --> 0.000 AND 0.5014 COEF
3) PROD_TECHNICAL_NEGATIVE --> 0.000 AND -0.4577 COEF
4) HEDONIC --> 0.009 AND 0.1952 COEF
5)UTILITARIAN --> 0.012 AND 0.1987 COEF
6) PROD_DESIGN_POSITIVE --> 0.007 AND 0.1792 COEF

Logit Regression Results			
Dep. Variable:	Purchase	No. Observations:	600
Model:	Logit	Df Residuals:	582
Method:	MLE	Df Model:	17
Date:	Wed, 20 Mar 2024	Pseudo R-squ.:	-0.8368
Time:	21:35:32	Log-Likelihood:	-341.84
converged:	False	LL-Null:	-186.11
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-12.4571	2.347	-5.307	0.000	-17.058	-7.856
Number_Words_Review	0.0060	0.005	1.255	0.209	-0.003	0.015
Prod_Design_positive	1.0683	0.405	2.635	0.008	0.274	1.863
Prod_Design_negative	-0.4830	0.543	-0.889	0.374	-1.547	0.581
Prod_Technical_positive	-0.3856	0.535	-0.721	0.471	-1.434	0.663
Prod_Technical_negative	1.2567	0.532	2.360	0.018	0.213	2.300
Prod_Price_positive	0.7553	0.410	1.843	0.065	-0.048	1.558
Prod_Price_negative	1.7412	1.020	1.706	0.088	-0.259	3.741
Serv_Delivery_positive	-0.8470	0.894	-0.947	0.344	-2.600	0.906
Serv_Delivery_negative	26.1594	1008.756	0.026	0.979	-1950.967	2003.286
Country	1.3936	0.399	3.491	0.000	0.611	2.176
Gender	0.4514	0.361	1.252	0.211	-0.255	1.158
Sentiment	0.6074	0.291	2.085	0.037	0.036	1.178
Rating_Score	1.1698	0.472	2.478	0.013	0.245	2.095
Number_of_Purchases	2.0177	0.281	7.175	0.000	1.467	2.569

ANSWER #2

- WE CHOOSE TO USE LOGISTIC REGRESSION BECAUSE THE DEPENDENT VARIABLE WHICH IS THE PURCHASE COLUMN HAVE BINARY OUTCOMES.
- WE USED LOGIT INSTEAD OF PROBIT BECAUSE THE DEPENDENT VARIABLE IS CONSIDERED TO BE A TRULY QUALITATIVE CHARACTER.

THESE ARE THE MAIN PREDICTORS OF A PURCHASE CONVERSION:

1) BEING AN EXISTING CUSTOMER (NR. OF PURCHASES)--> 0.000 AND 2.0177 AS COEF
2) COUNTRY --> 0.000 AND 1.3936 AS COEF
3) PROD_TECH_NEG --> 0.018 AND 1.2567 COEF
4) RATING SCORE --> 0.013 AND 1.1698 COEF
5)SENTIMENT --> 0.037 AND 0.6074 COEF

GEE Regression Results

Dep. Variable:	Number_of_Purchases	No. Observations:	600
Model:	GEE	No. clusters:	600
Method:	Generalized Estimating Equations	Min. cluster size:	1
		Max. cluster size:	1
Family:	Poisson	Mean cluster size:	1.0
Dependence structure:	Independence	Num. iterations:	1
Date:	Wed, 20 Mar 2024	Scale:	1.000
Covariance type:	robust	Time:	21:35:33

	coef	std err	z	P>|z|	[0.025	0.975]
Review_ID	0.0006	0.000	1.569	0.117	-0.000	0.001
Number_Words_Review	-0.0014	0.002	-0.882	0.378	-0.004	0.002
Prod_Design_positive	-0.2205	0.124	-1.779	0.075	-0.464	0.022
Prod_Design_negative	-0.3373	0.148	-2.284	0.022	-0.627	-0.048
Prod_Technical_positive	0.1390	0.176	0.788	0.431	-0.207	0.484
Prod_Technical_negative	-0.4124	0.172	-2.401	0.016	-0.749	-0.076
Prod_Price_positive	-0.2448	0.137	-1.787	0.074	-0.513	0.024
Prod_Price_negative	-0.0219	0.277	-0.079	0.937	-0.565	0.521
Serv_Delivery_positive	0.3449	0.250	1.380	0.168	-0.145	0.835
Serv_Delivery_negative	-0.8289	0.636	-1.304	0.192	-2.075	0.417
Country	-0.1284	0.145	-0.884	0.377	-0.413	0.156
Gender	-0.1555	0.115	-1.347	0.178	-0.382	0.071
Sentiment	0.0741	0.085	0.874	0.382	-0.092	0.240
Rating_Score	-0.1770	0.077	-2.293	0.022	-0.328	-0.026
Purchase	1.2492	0.125	9.960	0.000	1.003	1.495
Age_Range_36-65	-0.2323	0.113	-2.047	0.041	-0.455	-0.010
Utilitarian	-0.1762	0.146	-1.205	0.228	-0.463	0.110
Hedonic	-0.1410	0.151	-0.935	0.350	-0.437	0.155

Skew:	1.4574	Kurtosis:	1.8889

ANSWER #3

- HERE WE CHOOSE POISSON BECAUSE THE NUMBER OF TOTAL PURCHASES IS A COUNT VARIABLE.

- ACCORDING TO THE RESULTS OF GEE REGRESSION, SIGNIFICANT PREDICTORS ASSOCIATED WITH PRIOR PURCHASES ARE:

1) PURCHASE --> 0.000 AND 1.2492 COEF
2) PROD_TECH_NEG --> 0.016 AND -0.4124 COEF
3) PROD_DESIGN_NEG --> 0.022 AND -0.3373 COEF
4) AGE RANGE --> 0.041 AND -0.2323 COEF
5) RATING SCORE --> 0.022 AND -0.177 COEF

ANSWER #4.1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Rating_Score   R-squared:                       0.636
Model:                            OLS   Adj. R-squared:                  0.634
Method:                 Least Squares   F-statistic:                     347.6
Date:                Wed, 20 Mar 2024   Prob (F-statistic):          1.95e-130
Time:                        21:35:33   Log-Likelihood:                -676.27
No. Observations:                 600   AIC:                             1361.
Df Residuals:                     596   BIC:                             1378.
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                                0.9533      0.109      8.706      0.000       0.738       1.168
Prod_Design_positive                 2.0256      0.235      8.628      0.000       1.565       2.487
Sentiment                            0.8610      0.031     28.025      0.000       0.801       0.921
Product_Design_Sentiment_Interact   -0.4616      0.058     -7.961      0.000      -0.575      -0.348
==============================================================================
Omnibus:                       42.508   Durbin-Watson:                   1.842
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.610
Skew:                          -0.539   Prob(JB):                     2.54e-14
Kurtosis:                       4.158   Cond. No.                         38.3
==============================================================================
```

- The model aims to investigate the influence of product design ratings, sentiment scores, and their interaction on rating scores.
- Overall, the results suggest that while both product design ratings and sentiment scores individually have a positive impact on rating scores, their interaction has a negative impact, indicating a potential complexity in the relationship between these variables.

ANSWER #4.2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Rating_Score   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     10.12
Date:                Wed, 20 Mar 2024   Prob (F-statistic):           1.64e-06
Time:                        21:35:34   Log-Likelihood:                -964.80
No. Observations:                 600   AIC:                             1938.
Df Residuals:                     596   BIC:                             1955.
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const                    4.4902      0.120     37.411      0.000       4.254       4.726
Country                 -0.5076      0.165     -3.079      0.002      -0.831      -0.184
Age_Range_36-65         -0.0458      0.148     -0.310      0.757      -0.336       0.244
Country_Age_Interact    -0.0558      0.206     -0.270      0.787      -0.461       0.349
==============================================================================
Omnibus:                      118.805   Durbin-Watson:                   1.722
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              189.562
Skew:                          -1.314   Prob(JB):                     6.87e-42
Kurtosis:                       3.824   Cond. No.                         8.14
==============================================================================
```

- The model aims to investigate the influence of the country where the review was provided, the age range of the reviewer, and their interaction on rating scores.
- Additionally, the negative coefficients indicate that higher values of these variables are associated with lower rating scores.

CONCLUSION

- **A1:** Positive aspects of product design, technical aspects, sentiment and hedonic/utilitarian categories have a positive impact in Rating Score, while discontent with technical aspects and delivery service impact negatively. Positive impact meaning that customers tend to give a high rating score.
- **A2:** Positive aspects of product design, sentiment, rating score and hedonic category have a positive impact in purchasing a product, also if the person is from Belgium, is more likely to purchase. The marketing team should target more the Belgian market to increase sales. Technical issues in products are actually reducing effective purchases so they should focus in quality and post sales assistance.
- **A3:** We found that people between ages from 36 to 65 years are less likely to do several number of purchases. The marketing team should people between 18-35 years or adequate the product for adults needs. Technical and design issues discourages potential customers to buy more products, as also this customers are giving bad reviews to the products, decreasing the amount of sales.
- **A4.1:** The results suggest that both product design ratings and sentiment scores, as well as their interaction, have a significant impact on rating scores. Additionally, the interaction between product design ratings and sentiment scores further enhances the predictive power of the model, indicating that the combined effect of positive sentiment and positive product design ratings leads to higher rating scores.
- **A4.2:** The results suggest that the country where the review was provided significantly influences rating scores, while the age range of the reviewer and the interaction between country and age range do not have a significant impact on rating scores.

GROUP DS2

THANKS!