Czech University of Life Sciences Prague

Faculty of Economics and Management
Informatics

Statistical Data Analysis
Semestral Project

By: Tural Dadashzade
Usman Ullah Rehmat Ullah
AddiskiDAN Tegegn
Regeczy Mikulas

Instructors:
Ing. Tomas Hlavsa Ph.D.

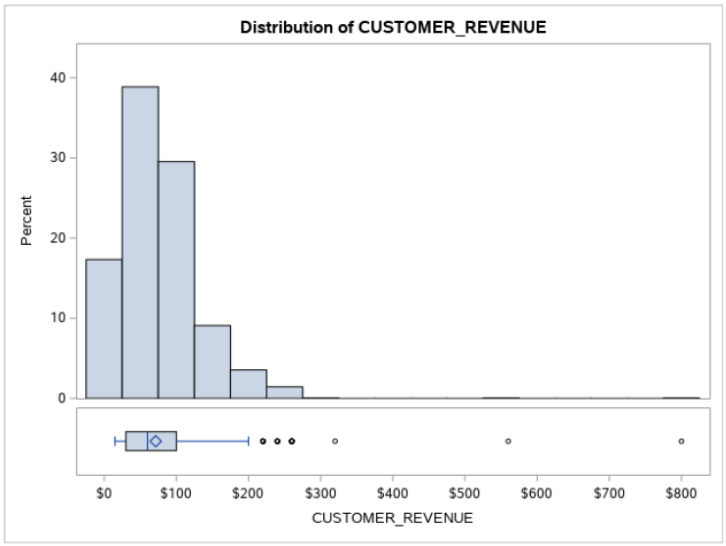
December 2021

Part 1

First, we have filtered the data, specified the Genders as Males and Females as per the requirement. Afterwards, we have obtained the summary results for the customer revenue.

11/10/21, 4:06 PM Results: Summary Statistics

Analysis Variable : CUSTOMER_REVENUE CUSTOMER_REVENUE									
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Lower Quartile	Upper Quartile
71.4169675	55.9900118	15.0000000	800.0000000	60.0000000	1662	3134.88	15.0000000	30.0000000	100.0000000



Later, we have created 2 categories with the obtained quartiles where the upper quartile was 100 and lower quartile was 30. After the creation of categories, we have created the contingency table with CHI squared value to test if we can accept the Null hypothesis. Taking into consideration that the Chi Squared value is higher than the critical value of 3.841 we can reject the NULL hypothesis which means there might be a relationship between Gender and Revenue.

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.7583	0.5992	0.9596
Relative Risk (Column 1)	0.9406	0.8918	0.9920
Relative Risk (Column 2)	1.2404	1.0334	1.4889

Sample Size = 1662

Frequency
Expected

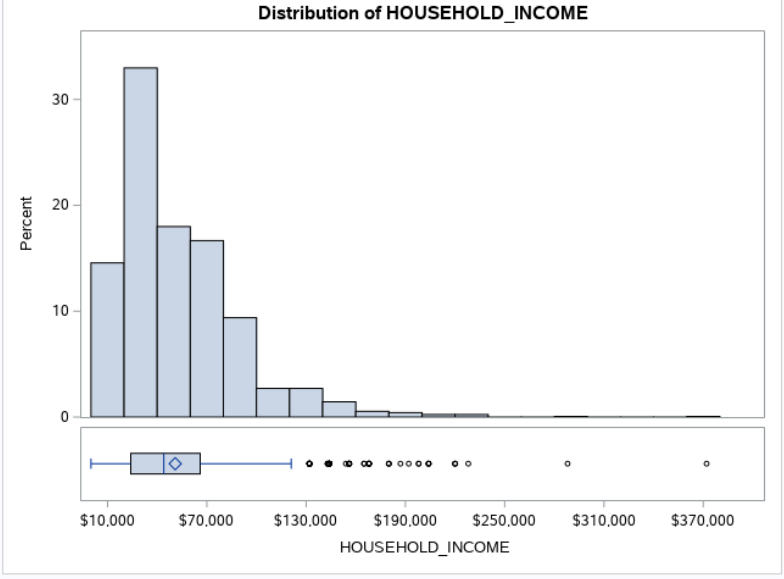
Table of GENDER by Revenue_Category			
	Revenue_Category		
GENDER(GENDER)	0	1	Total
Female	500 518.99	163 144.01	663
Male	801 782.01	198 216.99	999
Total	1301	361	1662

Frequency
Expected

Table of GENDER by Revenue_Category			
	Revenue_Category		
GENDER(GENDER)	0	1	Total
Female	500 518.99	163 144.01	663
Male	801 782.01	198 216.99	999
Total	1301	361	1662

Similarly, to the Customer Revenue, we have obtained **the data for Household income** to get its summary statistics initially.

Analysis Variable : HOUSEHOLD_INCOME HOUSEHOLD_INCOME									
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Lower Quartile	Upper Quartile
50891.10	36910.61	0	372000.00	44000.00	1662	1362393310	36000.00	24000.00	66000.00



Then we have created **4 different categories for Household income** based on the lower percentile, Median, and Upper Percentile. Then we did the Chi squared test in order to see if we can accept the Null hypothesis. As we can see Chi squared value is bigger than the critical value, so we reject the Null hypothesis. There is a relationship between Household income and Revenue.

Frequency
Expected

Table of House_cat by Revenue_Category			
	Revenue_Category		
House_cat	0	1	Total
1	452 396.88	55 110.12	507
2	310 291.98	63 81.019	373
3	287 288.85	82 80.15	369
4	252 323.29	161 89.707	413
Total	1301	361	1662

Statistic	DF	Value	Prob
Chi-Square	3	112.8044	<.0001
Likelihood Ratio Chi-Square	3	109.4520	<.0001
Mantel-Haenszel Chi-Square	1	104.3516	<.0001
Phi Coefficient		0.2605	
Contingency Coefficient		0.2521	
Cramer's V		0.2605	

Sample Size = 1662

Finding the residuals went through creating a suitable table for GENMOD procedure in SAS where we have created dependent variable C with using customer revenue category and household income category. Considering the fact that we had 4x2 CONTINGENCY table we had 8 observations.

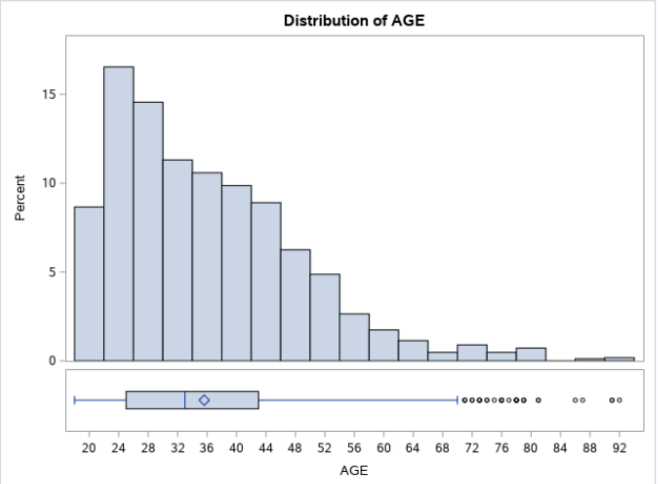
Observation Statistics						
Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	55.124549	2.7670559	2.7064455	6.9660359	7.1220391	7.0987105
2	18.018652	1.0544955	1.0439195	2.5434193	2.569187	2.5648643
3	-1.850181	-0.108862	-0.108979	-0.265107	-0.264823	-0.264871
4	-71.29302	-3.965052	-4.126078	-10.21254	-9.813977	-9.880133
5	-55.12455	-5.252946	-5.820518	-7.891563	-7.122039	-7.550392
6	-18.01865	-2.001842	-2.083909	-2.674512	-2.569187	-2.633634

Observation Statistics						
Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
7	1.8501805	0.206663	0.2058755	0.263814	0.2648231	0.264209
8	71.29302	7.5272077	6.7629523	8.8175407	9.8139773	9.2408235

Part 2

As we can see from the **descriptive data** of the age variable, we can see a range between 18 and 92 and also the average of 35 years.

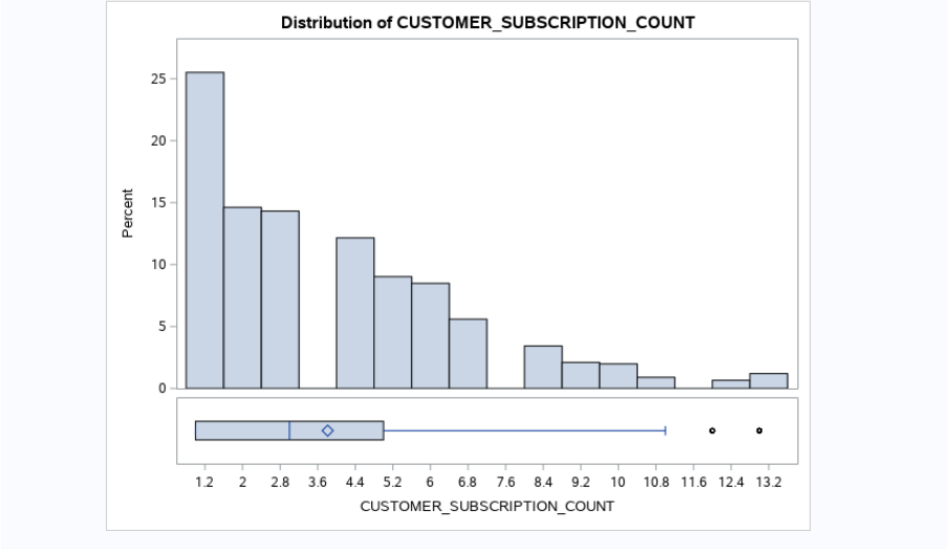
Analysis Variable : AGE AGE											
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Std Error	Variance	Mode	Range	Sum	Coeff of Variation
35.6161252	12.6733872	18.0000000	92.0000000	33.0000000	1662	0	0.3108688	160.6147428	21.0000000	74.0000000	59194.00	35.5832846



To summarize the informan for customer subscription count we can mention that the data is not bell shaped and shows the minimum of 1 subscription whereas maximum is 13.

Analysis Variable : CUSTOMER_SUBSCRIPTION_COUNT CUSTOMER_SUBSCRIPTION_COUNT

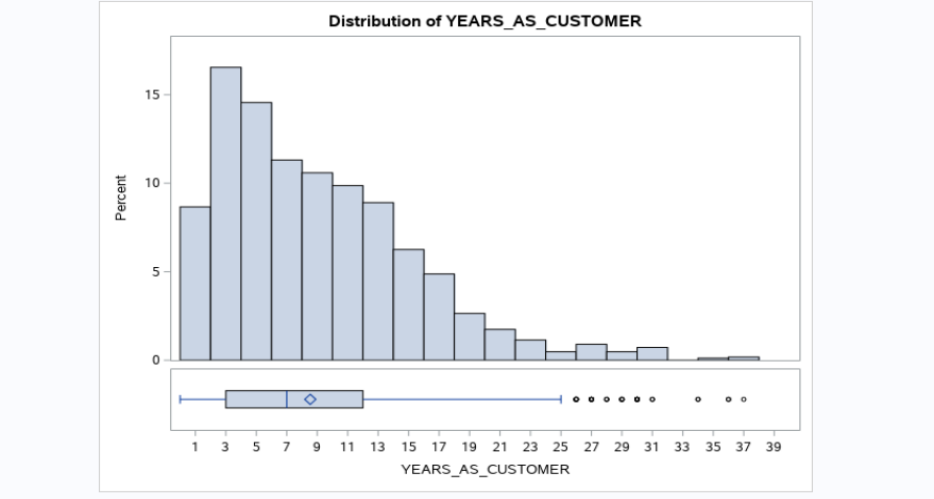
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Std Error	Variance	Mode	Range	Sum	Coeff of Variation
3.8128761	2.7490685	1.0000000	13.0000000	3.0000000	1662	0	0.0674326	7.5573778	1.0000000	12.0000000	6337.00	72.0996037



The most important factor in summary statistics for years as customers would be having the highest coefficient of variation which means the data is more spread out than the other 2 variables mentioned above.

Analysis Variable : YEARS_AS_CUSTOMER YEARS_AS_CUSTOMER

Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Std Error	Variance	Mode	Range	Sum	Coeff of Variation
8.5439230	6.3355069	0	37.0000000	7.0000000	1662	0	0.1554053	40.1386476	2.0000000	37.0000000	14200.00	74.1522004

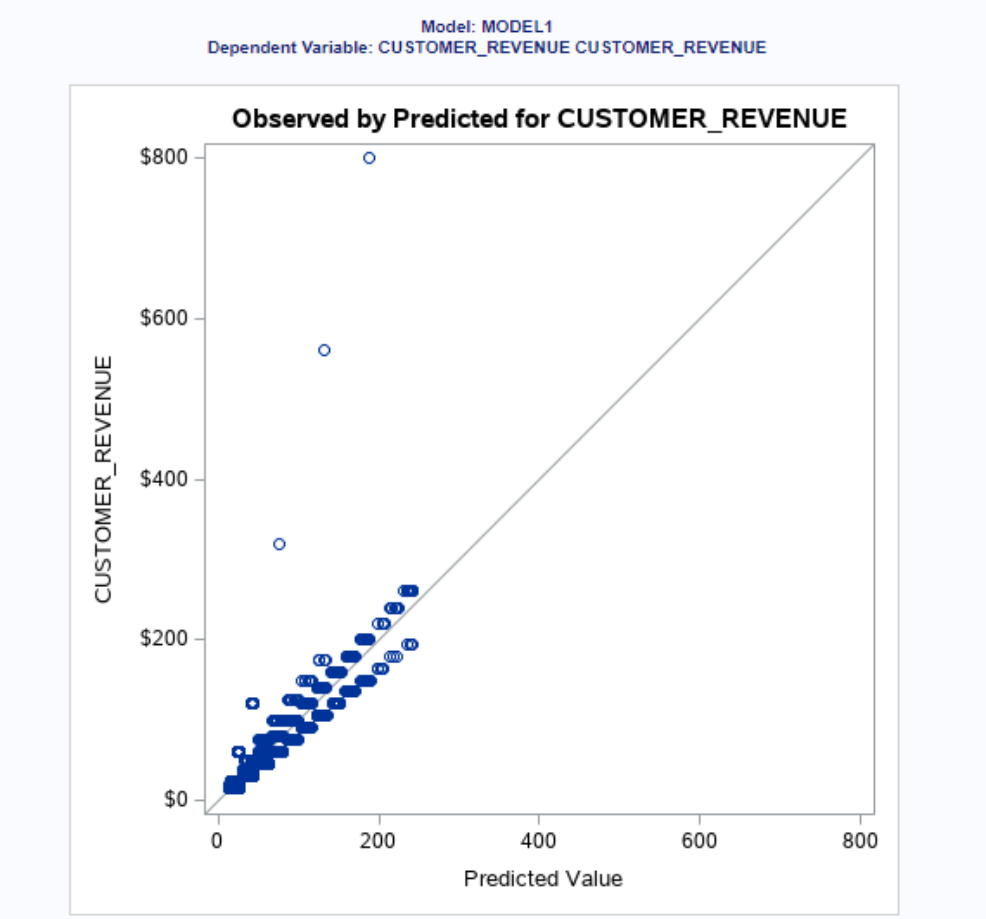


During the collinearity analysis before regression, we see that there is a high relationship between age and years as customer. Due to that we will perform the regression analysis without the age group.

4 Variables:	AGE HOUSEHOLD_INCOME YEARS_AS_CUSTOMER CUSTOMER_SUBSCRIPTION_COUNT

Pearson Correlation Coefficients, N = 1662				
	AGE	HOUSEHOLD_INCOME	YEARS_AS_CUSTOMER	CUSTOMER_SUBSCRIPTION_COUNT
AGE AGE	1.00000	0.14209	0.99922	-0.16592
HOUSEHOLD_INCOME HOUSEHOLD_INCOME	0.14209	1.00000	0.14357	0.35930
YEARS_AS_CUSTOMER YEARS_AS_CUSTOMER	0.99922	0.14357	1.00000	-0.16529
CUSTOMER_SUBSCRIPTION_COUNT CUSTOMER_SUBSCRIPTION_COUNT	-0.16592	0.35930	-0.16529	1.00000

This is the linear model for customer revenue based on 1662 observations:



From the below tables we can see that we have very small P value in ANOVA. It means the model we obtained is really doing a great job. As R squared varies between 0 and 1, 78 percent is good number in order to explain the fit of linear regression in our data. As the P values are very small for Household Income and Customer subscription count, we can say that our parameters are statistically significant for those 2. We also see that Variance Inflation is small for Household income and customer subscriptions.

Model: MODEL1
Dependent Variable: CUSTOMER_REVENUE CUSTOMER_REVENUE

Number of Observations Read	1662
Number of Observations Used	1662

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4092939	1364313	2030.37	<.0001
Error	1658	1114099	671.95364		
Corrected Total	1661	5207038			

Root MSE	25.92207	R-Square	0.7860
Dependent Mean	71.41697	Adj R-Sq	0.7857
Coeff Var	36.29679		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.86577	1.53828	0.56	0.5736	.	0
HOUSEHOLD_INCOME	HOUSEHOLD_INCOME	1	0.00006471	0.00001893	3.42	0.0006	0.82855	1.20693
YEARS_AS_CUSTOMER	YEARS_AS_CUSTOMER	1	-0.03339	0.10436	-0.32	0.7490	0.92538	1.08064
CUSTOMER_SUBSCRIPTION_COUNT	CUSTOMER_SUBSCRIPTION_COUNT	1	17.71456	0.25505	69.45	<.0001	0.82288	1.21525

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	HOUSEHOLD_INCOME	YEARS_AS_CUSTOMER	CUSTOMER_SUBSCRIPTION_COUNT
1	3.25425	1.00000	0.01460	0.02250	0.02271	0.02077
2	0.41685	2.79406	0.00183	0.02647	0.46086	0.25073
3	0.21160	3.92166	0.07636	0.95095	0.01690	0.25226
4	0.11730	5.26706	0.90721	0.00006935	0.49953	0.47624

PART 3:

As we have Age household income and customer subscription count as quantitative vairables, we first check the collinearity before building binary logistic regression. From the below **collinearity analysis** we will not need to eliminate any of the variables as the high relationship is not on the scene.

3 Variables:	AGE HOUSEHOLD_INCOME CUSTOMER_SUBSCRIPTION_COUNT

Pearson Correlation Coefficients, N = 1662

	AGE	HOUSEHOLD_INCOME	CUSTOMER_SUBSCRIPTION_COUNT
AGE AGE	1.00000	0.14209	-0.16592
HOUSEHOLD_INCOME HOUSEHOLD_INCOME	0.14209	1.00000	0.35930
CUSTOMER_SUBSCRIPTION_COUNT CUSTOMER_SUBSCRIPTION_COUNT	-0.16592	0.35930	1.00000

As you can see from the overview of **Binary logistic regression**, we have 1662 observations in total in which we have 1301 in lower revenue category, and 361 in higher revenue category

Model Information	
Data Set	WORK.NEWPER1
Response Variable	Revenue_Category
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1662
Number of Observations Used	1662

Response Profile		
Ordered Value	Revenue_Category	Total Frequency
1	0	1301
2	1	361

Probability modeled is Revenue_Category=1.

Class Level Information			
Class	Value	Design Variables	
GENDER	Female	1	0
	Male	0	1

From the Global Null hypothesis, we can see that the parameters are statistically significant which means at least one beta is different than 0 which will let us to reject Null hypothesis. When we look to the individual analysis, we can see that Customer subscription count is the beta which is statistically significant. Lastly, in the odds ratio we can see that Females tend to have 0.84 times of man to be in higher revenue category. Also, we can see that if you have more customer subscription counts then almost 4 times higher chance of being in top 25 percent revenue category.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1165.4948	4	<.0001
Score	962.4849	4	<.0001
Wald	261.4778	4	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GENDER	1	0.5725	0.4493
AGE	1	0.1742	0.6764
HOUSEHOLD_INCOME	1	0.0610	0.8049
CUSTOMER_SUBSCRIPTIO	1	250.3032	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-9.0648	0.6452	197.3650	<.0001
GENDER	Female	1	-0.1694	0.2239	0.5725	0.4493
GENDER	Male	0	0	.	.	.
AGE		1	0.00419	0.0100	0.1742	0.6764
HOUSEHOLD_INCOME		1	-7.15E-7	2.893E-6	0.0610	0.8049
CUSTOMER_SUBSCRIPTIO		1	1.5209	0.0961	250.3032	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GENDER Female vs Male	0.844	0.544	1.309
AGE	1.004	0.985	1.024
HOUSEHOLD_INCOME	1.000	1.000	1.000
CUSTOMER_SUBSCRIPTIO	4.576	3.790	5.525

And finally, in order to close our analysis, we will look to the **Confusion Metrics**. F_Revenue is representing the observed points and I_Revenue is representing the predicted points. Taking this into consideration we can say that 93 percent of our predictions for the revenue category was correct based on the confusion metrics

Frequency

Table of F_Revenue_Category by I_Revenue_Category

F_Revenue_Category(From: Revenue_Category)	I_Revenue_Category(Into: Revenue_Category) 0	1	Total
0	1255	46	1301
1	56	305	361
Total	1311	351	1662