DA 6223 Exam II

Due April 5, 2020

Reminder: Please follow the instructions below to complete this exam.

- The SAS Enterprise Guide project (.egp) file needs to be turned in **through Blackboard before 11:59 PM CT on the due day**. **No late** exam will be graded.
- The **SAS Enterprise Guide project (.egp) file needs to be formatted as Firstname_Lastname_ABC123_EXAM2.egp (For example, if a student is named Wenbo Wu, his SAS Enterprise Guide project file should be named Wenbo_Wu_ABC123_EXAM2.egp)**. **Failing to name the project file in the required way will result a 5% deduction on the grade.**
- You should **create separate Process Flows** for each problem. **Failing to do so will result a 10% deduction on the grade.**
- **For Problem 2**, please use the Note function in SAS Enterprise Guide to write down answers for each part.
- **All solutions must be written up independently, without looking at another student's work.**

For each student, the homework data files are saved under

O:\MSDA2020\DA6223_002\Instructor\Data\Exam 2\StudentID

where the StudentID is formatted as X_Y, for X to be the first name of the student, and Y to be the ABC123 ID number. For example, the StudentID for Wenbo Wu, will be Wenbo_ABC123.

Problem 1 (20 points): Transactions Data

In this problem, you will be working with the same **Demographics** data set that you have worked on for the midterm exam.

(a) Remove the "z_" part of the variable STATUS, EDU_LEVEL and AREA. Save the output SAS dataset Demographics_Modify in the WORK library. (8 Points; **Hint:** You can either recode the column or use the TRANWRD function in SAS)

(b) **Based on the Demographics_Modify data set** from last task, recode the following columns: (10 Points)

- Recode the GENDER column such that the recoded column take value "F" for females, "M" for males, and an empty space for unknown gender. Change the length of the recoded Gender to be 1.
- Recode the INCOME column such that the recoded column takes value "Low" for Income < \$19999.99; "Medium" for Income between \$20000 and \$69999.99; and "High" for Income >= \$70000.

- Recode the EDU_LEVEL (after removing the "z_" part from the previous task) column such that the recoded column takes value "Secondary" for "High School" and "<High School"; "College" for "Bachelors"; and "Graduate" for "Masters" and "PhD".

(c) Use the One-Way frequency table to summarize the recoded columns for GENDER, INCOME, STATUS, EDU_LEVEL, and AREA. (2 Points)

Problem 2 (55 points): Transactions Data

In order to plan innovative promotions to move items that are often purchased together, a store is interested in market basket analysis of items purchased from the Health and Beauty Aids Department and the Stationary Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. The **TRANSACTIONS** data set contains information about more than 300,000 transactions made over the past three months. Use appropriate procedure in SAS Enterprise Guide to complete the following tasks.

(a) Import the transaction data and describe the attributes of the data set including: number of observations, number of columns, column type. (5 Points)

(b) Use One-Way Frequency Task to find out across all the stores, which product has been sold most in terms of the **quantity**. (Hint: The total quantity is NOT how many times this product is shown up in the data.) (5 Points)

(c) Which transaction yields the largest total amount? What is the transaction ID for this transaction and what is the total amount spent? (5 Points)

(d) Which transaction has bought most different types of products? How many different types of products were bought?(5 Points)

(e) Find out which store generates the largest revenue. What is the total revenue generated? (7 Points)

(f) Use Recode Column task to define a **Department** column using the following information. (5 Points)

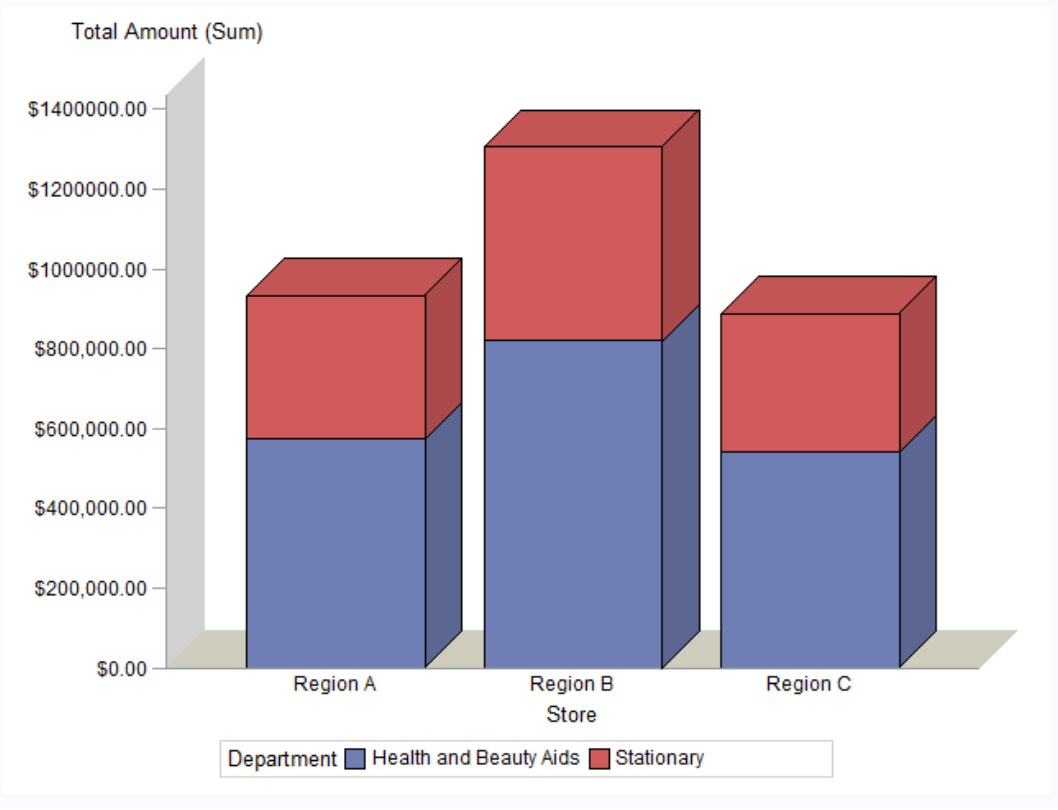
(g) The 10 stores in the data belong to different regions. Stores 1, 2, and 3 belong to Region A; stores 4, 5, 6, and 7 belong to Region B; and stores 8, 9, and 10 belong to Region C. Define a user-defined format for stores and use a Summary Tables task to generate the following output table. (9 Points)

(h) Construct the following Stacked Bar Chart (Hint: Applying the use-defined format). (7 Points)

(i) Applying a **Two Sample T-Test** to test whether the unit price of magazine is different between Store 1 and Store 10. (7 Points)

Product	Department
Candy Bar	Stationary
Deodorant	Health and Beauty Aids
Greeting Cards	Stationary
Magazine	Stationary
Markers	Stationary
Pain Reliever	Health and Beauty Aids
Pencils	Stationary
Pens	Stationary
Perfume	Health and Beauty Aids
Photo Processing	Stationary
Prescription Med	Health and Beauty Aids
Shampoo	Health and Beauty Aids
Soap	Health and Beauty Aids
Toothbrush	Health and Beauty Aids
Toothpaste	Health and Beauty Aids
Wrapping Paper	Stationary

	Store			Total
	Region A	Region B	Region C	
	Total Amount	Total Amount	Total Amount	Total Amount
Health and Beauty Aids	$573,990.77	$819,519.10	$541,554.84	$1935064.71
Stationary	$360,258.15	$486,939.76	$347,325.43	$1194523.33
Total	$934,248.92	$1306458.86	$888,880.27	$3129588.05



Problem 3 (25 points): Social Security Data

In this problem, you will be working with the **SocialSecurity.xlsx** data. Use appropriate procedure in SAS Enterprise Guide to complete the following tasks.

(a) Import the data in the **SS_Tax_Rate tab** of SocialSecurity.xlsx and describe the attributes of the data set including: number of observations, number of columns, column type. Read the problem statement to understand the meaning of each column. (3 Points)

(b) Take appropriate steps in SAS Enterprise Guide to reformat the imported data from the **SS_Tax_Rate** tab in a way that is presented in the following output. Please note that values and formats of the raw data in the SocialSecurity.xlsx **cannot** be changed. Please also pay attention to the variable type and format in the following output table. (16 Points)

	Begin	End	Employee_SSN_Tax_Rate	Employer_SSN_Tax_Rate	Total_SSN_Tax_Rate
1	1937	1949	1.00%	1.00%	2.00%
2	1950	1950	1.50%	1.50%	3.00%
3	1951	1953	1.50%	1.50%	3.00%
4	1954	1956	2.00%	2.00%	4.00%
5	1957	1958	2.25%	2.25%	4.50%
6	1959	1959	2.50%	2.50%	5.00%
7	1960	1961	3.00%	3.00%	6.00%
8	1962	1962	3.13%	3.13%	6.25%
9	1963	1965	3.63%	3.63%	7.25%
10	1966	1966	3.85%	3.85%	7.70%
11	1967	1967	3.90%	3.90%	7.80%
12	1968	1968	3.80%	3.80%	7.60%
13	1969	1969	4.20%	4.20%	8.40%
14	1970	1970	4.20%	4.20%	8.40%
15	1971	1972	4.60%	4.60%	9.20%
16	1973	1973	4.85%	4.85%	9.70%
17	1974	1977	4.95%	4.95%	9.90%
18	1978	1978	5.05%	5.05%	10.10%
19	1979	1979	5.08%	5.08%	10.16%
20	1980	1980	5.08%	5.08%	10.16%
21	1981	1981	5.35%	5.35%	10.70%
22	1982	1982	5.40%	5.40%	10.80%
23	1983	1983	5.40%	5.40%	10.80%
24	1984	1987	5.70%	5.70%	11.40%
25	1988	1989	6.06%	6.06%	12.12%
26	1990	1993	6.20%	6.20%	12.40%
27	1994	1996	6.20%	6.20%	12.40%
28	1997	1999	6.20%	6.20%	12.40%
29	2000	2015	6.20%	6.20%	12.40%

(c) Import the data in the **Wage_Limits** of SocialSecurity.xlsx. (3 Points)

(d) Merge the two imported data together so that the following output table can be obtained. (8 Points)

(e) Construct a scatter plot for above data. Using **SS_Wage_Limit** on the horizontal axis and **Employee_SSN_Tax_Rate** on the versicle axis. Fit a linear regression line going through the plotted points. (5 Points)

	Year	SS_Wage_Limit	Employee_SSN_Tax_Rate	Employer_SSN_Tax_Rate	Total_SSN_Tax_Rate
1	1972	9,000.00	4.60%	4.60%	9.20%
2	1973	10,800.00	4.85%	4.85%	9.70%
3	1974	13,200.00	4.95%	4.95%	9.90%
4	1975	14,100.00	4.95%	4.95%	9.90%
5	1976	15,300.00	4.95%	4.95%	9.90%
6	1977	16,500.00	4.95%	4.95%	9.90%
7	1978	17,700.00	5.05%	5.05%	10.10%
8	1979	22,900.00	5.08%	5.08%	10.16%
9	1980	25,900.00	5.08%	5.08%	10.16%
10	1981	29,700.00	5.35%	5.35%	10.70%
11	1982	32,400.00	5.40%	5.40%	10.80%
12	1983	35,700.00	5.40%	5.40%	10.80%
13	1984	37,800.00	5.70%	5.70%	11.40%
14	1985	39,600.00	5.70%	5.70%	11.40%
15	1986	42,000.00	5.70%	5.70%	11.40%
16	1987	43,800.00	5.70%	5.70%	11.40%
17	1988	45,000.00	6.06%	6.06%	12.12%
18	1989	48,000.00	6.06%	6.06%	12.12%
19	1990	51,300.00	6.20%	6.20%	12.40%
20	1991	53,400.00	6.20%	6.20%	12.40%
21	1992	55,500.00	6.20%	6.20%	12.40%
22	1993	57,600.00	6.20%	6.20%	12.40%
23	1994	60,600.00	6.20%	6.20%	12.40%
24	1995	61,200.00	6.20%	6.20%	12.40%
25	1996	62,700.00	6.20%	6.20%	12.40%
26	1997	65,400.00	6.20%	6.20%	12.40%
27	1998	68,400.00	6.20%	6.20%	12.40%
28	1999	72,600.00	6.20%	6.20%	12.40%
29	2000	76,200.00	6.20%	6.20%	12.40%
30	2001	80,400.00	6.20%	6.20%	12.40%
31	2002	84,900.00	6.20%	6.20%	12.40%
32	2003	87,000.00	6.20%	6.20%	12.40%
33	2004	87,900.00	6.20%	6.20%	12.40%
34	2005	90,000.00	6.20%	6.20%	12.40%
35	2006	94,200.00	6.20%	6.20%	12.40%
36	2007	97,500.00	6.20%	6.20%	12.40%
37	2008	102,000.00	6.20%	6.20%	12.40%
38	2009	106,800.00	6.20%	6.20%	12.40%
39	2010	106,800.00	6.20%	6.20%	12.40%
40	2011	106,800.00	6.20%	6.20%	12.40%
41	2012	110,100.00	6.20%	6.20%	12.40%
42	2013	113,700.00	6.20%	6.20%	12.40%
43	2014	117,000.00	6.20%	6.20%	12.40%