Контроль набутих знань

практика (тиждень 3)

Завдання

Для роботи використовується файли Titanic.csv, test.csv, train.csv і titanic3.csv.

Список обов'язкових кроків для виконання:

1. Ознайомитись з наборами даних.
2. Створити код в SAS Studio з назвою у фоматі - lastname_cw3.sas;
3. Створити макро-змінні: зі своїм SAS ID та зі своїм прізвищем.
   Далі ці макрозмінні необхідно використовувати в коді.
4. Створити бібліотеку(назва бібліотеки=прізвище інтерна), в якій будете зберігати лише фіналізовані датасети (проміжні датасети зберігайте в директорії WORK).
5. Файли train.csv і test.csv містять записи, в яких Name вказано з помилкою. (наприклад: Assaf Khalil, Mrs. Mariana (Miriam")" vs Assaf Khalil, Mrs. Mariana ("Miriam")). Програмно відкоригуйте Name так, щоб ці файли коректно змерджились з файлами Titanic.csv і titanic3.csv.
6. Створіть датасет **PASSENGER** об'єднавши датасети Titanic(усі змінні), test.csv і train.csv (PassengerId) і titanic3.csv (embarked, home_dest).
7. Створити нову змінну FareD як копію Fare, але з конвертацією у долари США (USD). У новій змінній FareD значення повинні відображатися у грошовому форматі зі знаком $, бути округленими до двох знаків після десяткової крапки.
8. За допомогою PROC FORMAT створіть формат і застосуйте до відповідних змінних:
   - Age – повинна мітити значення (Child, Teen, Adult). Значення для визначення цих категорій можете взяти з практичного завдання (тиждень 1);
   - Survived – повинна містити значення (Survived, Died);
   - Embarked – повинна містити значення (Cherbourg, Queenstown, Southampton);
   - Fare – повинна містити значення: Lower Quartile (перший квартиль), Median (другий та третій квартиль), Upper Quartile (четвертий квартиль). Перед створенням відповідного формату, за допомогою дата або процедурного кроку знайдіть ці значення.
9. Додати змінні Dmy - (дата відправки — 10 квітня 1912 для пасажирів з портів Cherbourg і Southampton, 11 квітня 1912 з порту Queenstown) у форматі date9; змінну BoardingDay - день тижня, на який припадає посадка кожного пасажира; змінну DaysAtSea, яка розраховує кількість днів, проведених у морі до вечора 14 квітня 1912, залежно від дати посадки пасажира.
10. Назва змінних і лейбли у фінальному датасеті (**PASSENGER**) мають відповідати Додатку 1. Якщо в додатку ці змінні(або аналоги) не представленні, то назвіть за власним бажанням.

11. Для змінної Fare у розрізі змінної CLASS необхідно обчислити наступні статистичні характеристики і побудувати звіт:
    - NOBS – кількість заповнених значень (тобто без пропусків);
    - MAX_Fare – максимальне значення;
    - MEAN_FARE – математичне сподівання;
    - CV_FARE – коефіцієнт варіації.

    Це потрібно зробити за допомогою двох різних способів:

    - Програма повинна бути реалізована **у вигляді кроку даних** (або декількох кроків даних), із використанням RETAIN та інших необхідних конструкцій. Отриманий датасет назвіть **FARE_data;**
    - За допомогою процедурного кроку використовуючи стандартні статистичні процедури, як то PROC MEANS/SUMMARY/FREQ/TABULATE. Отриманий датасет назвіть **FARE_proc.**

    Додаткове бонусне завдання: порівняйте отримані результати, які ви отримали (крок даних) із еталонним (процедурний крок), за допомогою процедури PROC COMPARE.

12. За допомогою PROC PRINT створіть репорт із отриманого датасету **FARE_data** (див.приклад).
    - На процедурному кроці PROC PRINT на змінну Class наложіть формат відповідно до Додатку 1.
    - При необхідності числа заукругліть за допомогою функції ROUND (в зразку зазначенно скільки цифр має бути після крапки. Наприклад: змінна MAX_Fare = xx.xx. Це означає що заукруглюємо до сотих.)

Прізвище ім’я інтерна

Заголовок звіту

Class	NOBS	MAX_Fare	Mean_Fare	CV_Fare
Upper	xxx	xx.xx	xx.xxx	xx.xxxx
Middle	xxx	xx.xx	xx.xxx	xx.xxxx
Lower	xxx	xx.xx	xx.xxx	xx.xxxx

Назва коду, який створює звіт Дата створення звіту у форматі DD-MM-YYYY HH:MM

13. Створіть звіт REPORT_lastname.rtf відповідно до прикладу (Додаток 2).

Загалом, можливі деякі моменти, що не досить формалізовані в постановці завдання, в цьому випадку раджу проявити ініціативу та креативність, яку опишіть у вигляді коментаря на початку вашої програми.

Додаток 1. Теоретична інформація.

VARIABLE DESCRIPTIONS

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Survival	Survival (0 = No; 1 = Yes)
name	Name
Sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
Home_dest	Home/Destination

SPECIAL NOTES

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)

If the Age is estimated, it is in the form xx.5

Fare is in Pre-1970 British Pounds ()

Conversion Factors: 1 = 12s = 240d and 1s = 20d

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic

Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations

Додаток 2. Приклад звіту

Прізвище ім'я інтерна

Дата побудови звіту у форматі DATETIMEw.

Заголовок звіту

Таблиця 1. Співвідношення загиблих/живих в залежності від статі

Стать			
Чоловіча		Жіноча	
Загинули	Вижили	Загинули	Вижили
Count	Count	Count	Count
Pct	Pct	Pct	Pct

Note: Count – кількість, Pct – відсоток (наприклад, жінок що вижили, серед усіх жінок).

Таблиця 2 Шанси на виживання, в залежності від віку та класу каюти

Клас каюти	Демографічна категорія	Кількість що вижили	Кількість що загинули	Відношення тих, що вижили, до загиблих
Перший	Діти, підлітки			
	Дорослі			
Другий	Діти, підлітки			
	Дорослі			
Третій	Діти, підлітки			
	Дорослі			

Note: додайте примітку, за якою формулою визначили відношення (остання колонка).

Бонусне завдання: Мінімальне значення в стовпчику "Відношення тих, що вижили, до загиблих" забарвити червоним кольором.