11 - Assignment (due Monday October 26)

Import the Kaggle dataset of customer credit and spending from the URL `https://manctsui.github.io/datasets/credit.csv` into SAS and call this `credit`.

Problem 1A

The column `default_status` has values 0, 1, and "ERROR". Fill in the ??? portions of the following code to obtain a new dataset called `credit_by_default_status` which omits rows where the `default_status` entry is "ERROR", and whose columns are `income` and `default_status`.

```
data credit_by_default_status;
    set credit;
    where ??????
    keep ??????
run;
```

Problem 1B

Use SAS to run a t-test to test the claim that there is no difference in average income between the groups which have `default_status` of 1 versus those of 0 at the $\alpha = 0.01$ significance level.

(a) Show the SAS code.
(b) State the hypotheses H_0 and H_a.
(c) Check assumptions to see if you can run a t-test.
(d) Assuming (c) is checked and OK, report the t-statistics you obtained and the conclusion to make (reject H_0 vs fail to reject H_0).

Problem 2

Determine which numeric variable in the dataset `credit` is most correlated with income. Report the SAS code as well as the answer.

Problem 3

Use a simple linear regression to fit `income` against `average_monthly_spend`.

(a) Show your SAS code.
(b) Report SST, SSM, SSE, and R^2 for your linear model. What % of variation is explained by your linear model?
(c) Comment on goodness of fit of your linear model.

Problem 4A

Consider the following synthetic data read into SAS as `IncomeExp`.

```
data IncomeExp;
   input Income Expenditure @@;
   datalines;
15 14   18 16   22 19   25 21   28 23   32 25
35 28   40 31   45 33   50 36   55 38   60 41
70 45   80 50   90 54   100 58  120 63  140 68
160 72  180 76  200 80  250 88  300 94  350 100
400 105 500 115 600 122 750 132 900 140 1000 148
;
run;
```

Run the following and comment on how appropriate a linear model is for fitting the data

```
proc reg data=IncomeExp;
   model Expenditure = Income;
    plot Expenditure*Income / NOSTAT;
run;
```

Problem 4B

```
data IncomeExp_Log;
   set IncomeExp;
   log_Income = log(Income);
   log_Exp = log(Expenditure);
run;

proc reg data=IncomeExp_Log;
   model log_Exp = log_Income;
    plot log_Exp*log_Income / NOSTAT;
run;
```

(a) Run the above and see how well `log_Exp` and `log_Income` are related by a linear model.
(b) Write the linear model out as an equation in terms of `Expenditure` and `Income` without the use of logarithms.
(c) Report SST, SSM, SSE, and R^2 for your linear model.