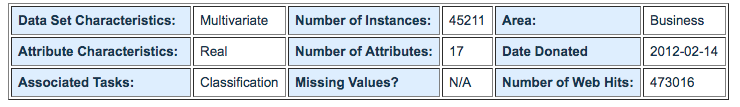
**Hands on challenge 6 - Regression**

Challenge 1

* Bank Marketing Data Set
  + Create a R markdown file to answer the questions below



**Attribute Information:**

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)   
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)   
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)  
  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

* **Questions:**

1. By using the fit function on a linear model, interpret the results
   1. What are the residuals (copy & paste)?

Residuals:

[ANSWER HERE]

- residuals are from zero or the average distance between the observed values and the model predictors

* 1. What are the coefficients?
     + Identify the strength of the coefficient to the dependent variables
  2. What is the residual standard error ?

[ANSWER HERE]

It's a measure of how close the fit is to the points. In this case 0.2 is fairly good at measuring prediction.

* 1. What is the R-Squared

[ANSWER HERE]

This is the proportion of the variance in the data that's explained by the model. The more variables you add - even if they don't help - the larger this will be.

What is the F-statistic?

[ANSWER HERE]

This is telling you whether the regression as a whole is performing 'better than random' - any set of random predictors will have some relationship with the response, so it's seeing whether your model fits better than you'd expect if all your predictors had no relationship with the response. This is used for a test of whether the model outperforms 'noise' as a predictor.

* 1. What is the p-value

[ANSWER HERE]

Most statisticians refer to statistically significant as P < 0.05 and statistically highly significant as P < 0.001 (less than one in a thousand chance of being wrong).

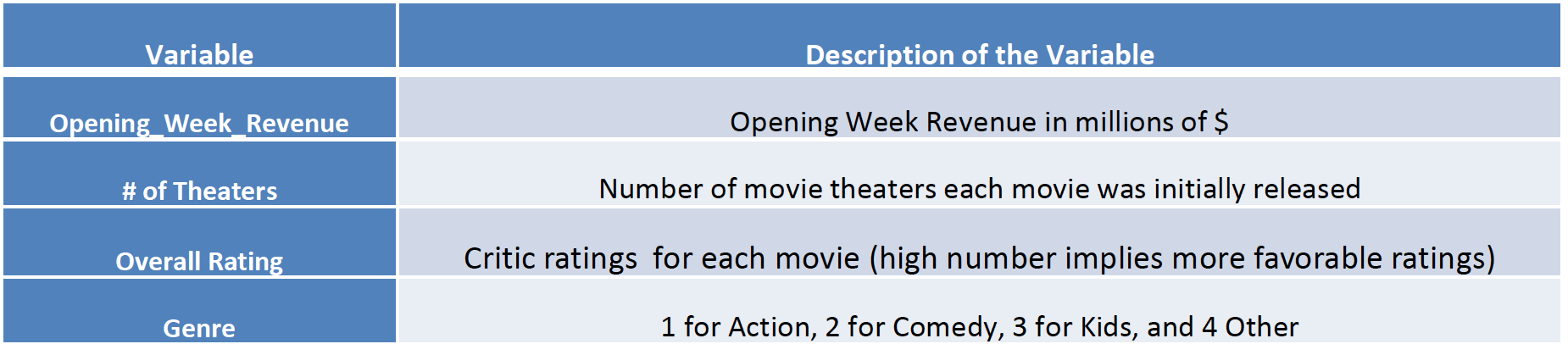
1. Which coefficients would you keep?
   1. Explain
2. Which coefficients would you remove?
   1. Explain
3. Run a cor plot against the coefficients and display the out put
4. How can we improve the business given the results from the strength in correlation from the coefficients (variables)?
   1. Explain which variables would lead to an understanding of how could we generate more profit?
   2. Explain which variables would lead to an understanding of which customers should we target?

**Challenge 2**

Linear Regression Movie Box Office Prediction

**Objective:**

Suppose you are helping Warner Bros. in developing a model for forecasting Box Office revenues for their new movie The Watchman. In the file “movie.csv” you are provided the opening week revenues (in millions of $) for various past movies along with several predictor variables:



Assignment

1. Develop a regression model for “Opening week Revenues” and all other variables as predictors. Interpret your parameters.
2. Prediction: The attributes for the movie “Watchman” are as follows:

–Theaters= 3611, Rating= 57, Genre= 1

Given this information, what are the predicted first week revenues for the new movie Watchman?