Regression and Multiple Regression

You work for a record company and your boss was interested in predicting album sales from advertising. You are provided with the data in a file.

There are 200 rows, each one representing a different album. There are also four columns. The first representing the sales of each album (in thousands) in the week after release. The second representing the amount (in thousands $) spent promoting the album before release. The third represents how many times it played on the Radio. The fourth column represents the attractiveness of the band.

Visualization

Visualizing data is important to get a 'feel' of what could be the relationship between variables be. For this part, please **draw the following scatterplots** between these variables

1. Sales and advertising
2. Sales and airplay
3. Sales and attractiveness

Please take screenshots for each of the above plots and include them in your Word file submission.

Linear Regression

4. Conduct a linear regression to **construct a linear model** between Sales and adverts and write down the **F-statistic and P-value**.
5. Discuss what these values (F-statistic and P-value) describe about our linear regression model? Is it good? Bad? Can't say?

Model Coefficients

6. What is the **intercept** value and **coefficient** (adverts) value of your linear regression model?

Knowing that the regression line is described using this equation
Y = $b_0 + b_1X_1$
Where,

Y denotes the sales
b_0 denotes the intercept value
b_1 denotes the advertising budget coefficient
X_1 denotes the advertising budget

The equation becomes:

Album sales = intercept value + (coefficient * advertising budget)

7. Using the intercept value and coefficient of your linear model, please **calculate how many records will be sold if we spent $135 000** on advertising the latest album "Dear Agony" by Breaking Benjamin

Multiple Regression

8. Conduct a multiple regression to construct a model between Sales and the predictors (adverts, airplay, attract) and report the **F-statistic and P-value.**

9. We know that the **R-squared value** can be used to evaluate the overall fit of a linear model. Also that **higher R-squared values are better if their p-values is < 0.05**.

   Based on this, discuss **which one of the two models that you constructed is better**?

   - Model 1: the linear model between Sales and advertising you constructed in (4).
   - Model 2: the multiple regression model between outcome: Sales and the predictors (advertising, airplay, attractiveness) that you constructed in (8).

Turn ins:

1. Your calculations, plots, screenshots, and your answers to the questions in a separate Word file.

2. The script (R or Python) that was used to load, correct, and impute the dataset.