Submit a report on a pdf with white background.

1. (20 pts.) Lets use `Cars93.csv` to clarify the relation between pivot tables and regression.

   a) Fit the model `m1 = smf.ols(formula = 'Price~C(Origin)',data = df).fit()`, then predict the price of a non-USA car. Finally, report a pivot table for the average Price by Origin.

   b) Fit the model `m2 = smf.ols(formula = 'Price~C(DriveTrain)*C(Origin)',data = df).fit()` then predict the price of a non-USA car with Front wheel drive. Report a pivot table for the average Price by Origin and DriveTrain. Does the predicted price agrees with the pivot table?

2. (40 pts.) Consider the `Hitters.csv` file. It is of interest to predict a player's salary based on the player's performance. Remove all rows with missing values. Split the data set into a training and test set (50%), use `random_state = 0`. Create an array of 100 α-values in the interval $10^{-2} < \alpha < 10^{10}$.

   a) The following loop fits ridge regression models with the α-values.

```
model = Ridge(normalize = True)
mspes = []
for i in alphas:
    model.set_params(alpha = i)
    model.fit(X_train, y_train)
    test_mspe = mean_squared_error(y_test, model.predict(X_test))
    mspes.append(test_mspe)
```

   Run the loop. For each model the `test_mspe` is found and stored into a list called `mspes`. Plot the `test_mspe` values (on y-axis) with alpha values on the (log) x-axis. Find the value of alpha minimizing the `test_mspe`.

   b) Fit a linear regression model and find the `test_mspe`. Add to the previous plot a horizontal dashed line at $y =$ `test_mspe` of the linear regression. Identify the set of alpha values that result in a ridge regression model with smaller `test_mspe` than the linear regression. Report the set of alpha values as an interval.

3. Split rows in `dataset.csv` into training (40%) and test set (use `stratify = y, random_state = 0`).

   a) (10 pts.) Fit a KNN model to predict y. Use the train and test sets to find the best number of neighbors in the range 1 to 20 (do not use K-fold cv). Find the test accuracy rate.

   b) (10 pts.) Fit a logistic regression model using predictors x_1 and x_2. Find test accuracy rate.

   c) (20 pts.) Use the train set to make a scatterplot of X_1 (x-axis) vs X_2, (y-axis). Use red color for rows with `y=1`. It should be clear that a linear boundary is not useful. So fit a (second order) logistic regression model to the train data with x_1^2, x_2^2, and $x_1 x_2$ as additional predictors. Find the test accuracy rate.